1. Task 1 – Exploratory Data Analysis

1.1. Identifiable Trends

Identifiable trends include justifiably high correlations between some features such as 'pclass' and 'fare' or 'sibsp' and 'parch'. This is explained by better classes requiring more expensive tickets and the Titanic mainly facilitated the immigration of third-class passengers [1]. This led to the inclusion of large families, having multiple parents/children and siblings/spouses. Other highly correlated features are discussed in Section 1.2.

Key observations include the disparity between male and female passengers and their respective survival rates. While men outnumber women by nearly 2-to-1, survival rates for women are much higher. Using only the bar chart given, it can be estimated that the female survival rate is approximately 80%, in comparison to men's 20%. As sex as a feature is only shown as a bar chart, its correlation with other features besides survivability is unknown.

1.2. Most and Least Important Features

As referenced in Section 1.1, the most important feature is expected to be sex, with pclass and fare also having high importance. This is evidenced by the near 60% in survivability rates by sex (bar chart), along with high correlation coefficients (-0.3 and 0.25 respectively). As women and wealthy passengers were given evacuation priority [2], it is quite certain that these features are most important when predicting survival.

The least important features can identified as 'parch', 'age', and 'sibsp', listed in order of descending importance. These features have correlation coefficients that are magnitudes lower than 'pclass' and 'fare', and do not have simple justifications for their relevance.

While 'sex' has been identified as highly important, there is uncertainty in the feature importance ranking. As sex has been removed from the feature correlation matrix, its quantitative score in comparison to other features has been lost. However, given the large difference in sex-based survivability, sex has been identified as the "most important" feature.

1.3. Possible Extensions in Exploratory Analysis

First, the correlation between sex and survival can be explored. Although sex is categorical, correlation tests such as Phi or point-biserial coefficients can be used to measure the correlation between two binary variables.

Second, further statistics such as the mean, variance, and size of categories can be calculated. If large class imbalances exist, they should be addressed during the training process. Once identified, up/down sampling or reweighting can be applied accordingly.

Lastly, more advanced feature analysis can be performed, such as principal component analysis (PCA) or feature importance ranking using random forests or gradient boosting machines. These approaches can identify feature importance past just statistical correlation and provide insights into feature engineering strategies.

2. Task 2 – Feature Attribution Explanations

2.1. Implementation of SHAP

2.1.1 Most and Least Important Features

While attribution scores are sample-specific, the average and standard deviation can generalise feature importance across all samples, as shown in Figure 1. For sample-specific results, two plots are shown in Appendix A.

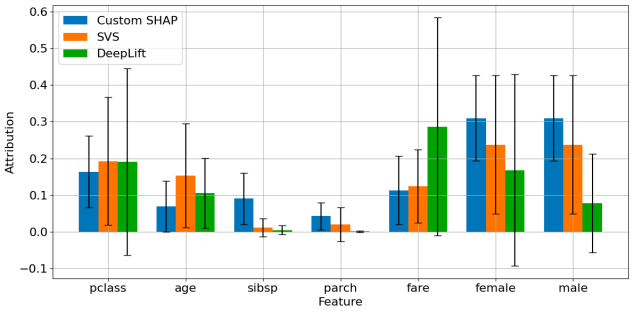

Figure 1: Mean and standard deviation of feature attribution scores across 10 Titanic Dataset samples, varying by method.

For the most important features, the SHAP methods similarly identify 'sex' and 'pclass'. While DeepLift identifies 'pclass' as well, it scores sex less significantly, instead identifying 'fare'.

For the least important features, all methods generally identify 'age', 'sibsp', and 'parch'. Notably, Deeplift also identifies 'male' while the SHAP methods identify 'fare' as potentially weak features.

2.1.2 Differences in SHAP Methods

As mentioned in 2.1.1, the SHAP methods rank most features similarly, only having small random variations in the attribution scores. These random variations are likely due to SVS's stochastic sampling of coalitions, unlike the custom method that considers every coalition possible. While more efficient, this sampling can introduce noise that reduces or increases attribution scores.

Conversely, Deeplift is substantially different to the SHAP methods, highly attributing 'fare' as the most important

feature while giving medium to low importance to ‘sex’. It also has higher standard deviation values and different attribution scores for ‘male’/‘female’. These differences are likely due to its unique methodologies, with the SHAP methods sharing a holistic approach that considers the effect of adding/removing features. Conversely, DeepLift calculates contributions based on the difference in activation of each neuron from a reference state.

2.1.3 Attribution Scores vs Expectations

For SHAP methods, attributions scores match expectations, ranking ‘sex’, then ‘fare and ‘pclass, and finally ‘age’, ‘sibsp’, and ‘parch’ in order of importance. While DeepLift’s least important features also match, it instead identifies ‘fare’ and ‘pclass’ as the most important.

The divergence between a user’s expectations and computed attributions is heavily dependent on the attribution methodology, such as DeepLift’s ability to highlight complex non-linear relationships between features. While accurate in the context of the model's learned parameters, might not align with initial expectations or simpler understandings of feature relationships. Conversely, the SHAP methods provide a balanced approach, considering many (if not all) possible feature interactions through its generated coalitions, which may align better with intuitive expectations.

2.1.4 Advantages and Disadvantages of SHAP Methods

Advantages of the SHAP methods include a model-agnostic approach that provides a more holistic, averaged, and intuitive view of feature importance [3]. Disadvantages include a reduced ability to capture complex feature relationships and increased computational strain when considering many features.

When considering the ‘custom’ method specifically, the advantages/disadvantages of SHAP are exacerbated, considering every possible coalition to provide more representative average scores at the cost of higher computational requirements. Conversely, SVS randomly samples coalitions, reducing computation times at the expense of scores being more vulnerable to random noise.

In contrast, DeepLift has greater computational efficiency by not needing to consider all possible coalitions. It can also capture complex non-linear relationships, providing detailed insights into how individual features and their changes from a reference point [4]. However, disadvantages include a model-specific approach that is heavily reliant on a reference state, possibly giving misleading results if not chosen correctly. Since the Titanic model is a shallow network and a default reference state was assigned, DeepLift’s advantages were not realised, and its scores did not match expectations.

2.2. Evaluation of Different Methods via Infidelity

To calculate infidelity, the perturb function’s standard deviation and resampling probability were investigated. Tables with exact results and a figure displaying infidelity vs noise scale can be found in Appendix B.

Table 1: Perturb Function’s Hyperparameter Search Space

	Standard deviation	Resample Probability
Values	[0.1, 0.5, 1, 2]	[0.2, 0.5, 0.8]

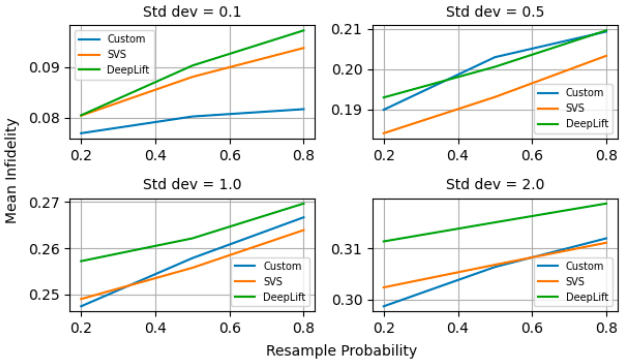

Figure 2: Mean infidelity vs combinations of standard deviation and resample probability

As seen in Figure 2, the DeepLift method consistently yields higher infidelity scores in comparison to SVS. DeepLift also surpasses the custom method for most cases, thus indicating its relatively lower performance among the three methods assessed.

When comparing SHAP methods, relative performance varies depending on the standard deviation values, as seen in Figure 2. Across the four standard deviations investigated, this variance is mostly symmetric, suggesting that the custom and SVS methods can regarded as functionally equivalent in terms of average infidelity.

2.3. Computational Efficiency of Different Methods

2.3.1 Key Performance Metrics for Dry Bean Model

Using an 80/20 train-test split, the Dry Bean model was trained and then evaluated with KPIs shown in Table 2.

Table 2: Dry Bean Model KPIs from Test Set

KPI	Accuracy	F1 score	Area Under the Curve
Value	0.93	0.93	0.81

2.3.2 Runtimes for Different Methods

All three methods were applied to the first 200 samples for both the Titanic and Dry Bean datasets.

Table 3: Runtime (Seconds) of Different Attribution Methods for the First 200 Samples in a Dataset

	Custom	SVS	DeepLift
Titanic	10.04	0.05	**0.02**
Dry Bean	19250.42	0.19	**0.02**

The most efficient method is DeepLift, executing the fastest for both the Titanic and Dry Bean datasets due to DeepLift's method of calculating neural activation in one forward and backward pass. Hence, the larger feature size of the Dry Bean dataset did not affect DeepLift's runtime.

By contrast, the SHAP methods become more computationally expensive as the feature space increases, leading to longer runtimes for both the custom method and SVS. While SVS samples only a fraction of coalition subsets, the subset diversity and complexity grow with feature size. Hence, getting a more representative sample of all possible combinations is also more complex.

However, sampling greatly reduces runtimes by reducing the necessary search space, unlike in the custom method. For every feature, every possible coalition is considered, resulting in a time complexity of 2^N where N is the feature size. Furthermore, the number of features for which to calculate an attribution score also increases. This is evidenced by the 5.3-hour runtime on Dry Bean's dataset.

3. Task 3 – Counterfactual Explanations

3.1. Distance Metrics

3.1.1 Standard vs Normalised L1 Distance

Standard L1 distance is a measure of the absolute differences between the coordinates of two points. This raw measure provides a simple interpretation of how much each feature must be altered to achieve the counterfactual.

Conversely, normalised L1 uses the relative distance between sample coordinates by considering the scale of each feature. This ensures that the distance measure is not dominated by features with inherently larger scales. As a result, normalised L1 provides a more balanced view of feature contributions irrespective of their original scales.

For counterfactuals, using normalized L1 can lead to more fair insights, especially in cases where features vary in scale and interpretability [5]. Furthermore, normalized L1 can produce counterfactuals that are potentially more actionable, as they consider the relative effort required to change each feature within its operational context.

3.1.2 Proposed Distance Function

As the pre-processed dataset is already inner quartile normalised, the proposed distance function can simply be the standard L1 distance. The inner quartile is also better at handling outliers than the suggested range-scaled normalisation method, treating each original feature equally while ensuring the variance of non-outlier points is not overly compressed.

3.2. NNCE and WAC Evaluation Metrics

After tuning WAC hyperparameters, results for λ=100 and lr = 0.00005 were selected to best showcase the difference between the NNCE and WAC regarding the three metrics.

Table 4: Validity, Proximity, and Plausibility Metrics for NNCE and WAC across 5 Trials of 20 Random Samples

Method	Validity		Proximity		Plausibility	
	μ	σ	μ	σ	μ	σ
NNCE	**1.0**	**0.0**	0.58	0.05	**0.15**	**0.04**
WAC	0.88	0.07	**0.28**	**0.03**	0.421	0.05

3.3. Differences between NNCE and WAC

As seen in Table 4, NNCE ensured perfect validity due to its selection of existing dataset instances, preconditioning every instance's validity before selecting the 'nearest' counterfactual. This consideration of only existing instances also makes NNCE model agnostic. Furthermore, these counterfactuals are more likely to be 'on manifold', increasing plausibility and ensuring that explanations are both understandable and realistic. However, this reliance on the existing dataset limits its ability to generate novel or closer counterfactuals, reducing its effectiveness in datasets with limited diversity. NNCE is also inherently inflexible as it does not have hyperparameters to tune.

Conversely, WAC offers a customizable and flexible approach that allows for closer counterfactuals to be found [6], evidenced by its lower proximity scores. By increasing λ and reducing the learning rate, the optimisation process heavily weighs the distance factor and makes smaller, more deliberate steps when finding 'nearby' counterfactuals. This results in counterfactuals nearer to the original instances, albeit potentially at the expense of increased computational time and uncertainty of finding valid counterfactuals.

While lower learning rates also improve plausibility, it is unlikely that WAC-generated counterfactuals will outperform NNCE. As seen in Table 4, plausibility scores improved to only 0.421 before validity dropped below 0.9, suggesting that comparable plausibility scores cannot be achieved while maintaining acceptable validity rates.

Furthermore, WAC does not ensure validity [7] as it creates new instances through an optimisation process that balances proximity and validity, controlled by λ. This may result in convergence to a theoretically optimal but practically invalid solution, evidenced by its imperfect validity score. WAC also depends on model differentiability, potentially limiting its applicability with non-differentiable models or very deep neural networks.

4. References

[1] National Oceanic and Atmospheric Administration. (2022, October 4). R.M.S Titanic - History and Significance | National Oceanic and Atmospheric Administration. https://www.noaa.gov/gc-international-section/rms-titanic-history-and-significance

[2] Bodowin College. (n.d.). Disproportionate Devastation | Titanic. Retrieved February 20, 2024, from https://courses.bowdoin.edu/history-2203-fall-2020-kmoyniha/reflection/

[3] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 2017-December, 4766–4775. https://arxiv.org/abs/1705.07874v2

[4] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. 34th International Conference on Machine Learning, ICML 2017, 7, 4844–4866. https://arxiv.org/abs/1704.02685v2

[5] Ye, Q., Li, Z., Fu, L., Zhang, Z., Yang, W., & Yang, G. (2019). Nonpeaked Discriminant Analysis for Data Representation. IEEE Transactions on Neural Networks and Learning Systems, 30(12), 3818–3832. https://doi.org/10.1109/TNNLS.2019.2944869

[6] Shao, X., Kersting, K., & Darmstadt, T. U. (2022). Gradient-based Counterfactual Explanations using Tractable Probabilistic Models. https://arxiv.org/abs/2205.07774v1

[7] Keane, M. T., & Smyth, B. (2020). Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12311 LNAI, 163–178. https://doi.org/10.1007/978-3-030-58342-2_11

5. Appendices

5.1. Appendix A

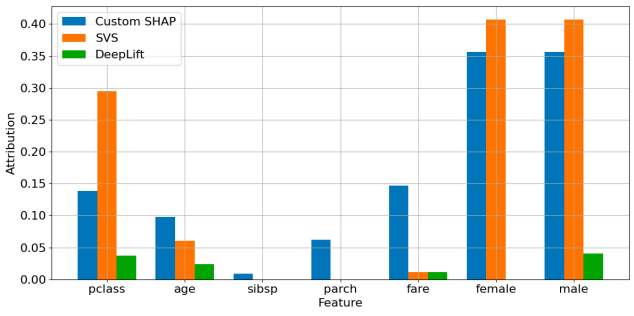

Figure 3: Feature attribution for random sample 1 in the Titanic dataset, varying by method.

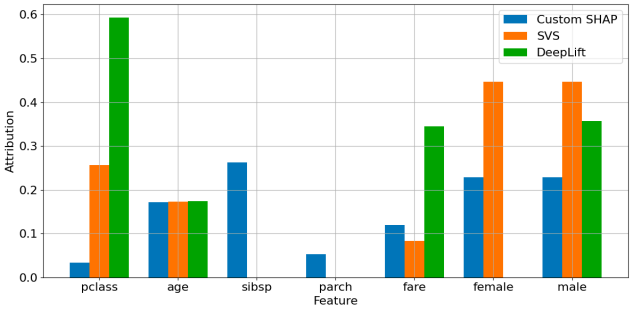

Figure 4: Feature attribution for random sample 2 in the Titanic dataset, varying by method.

5.2. Appendix B

Table 5: Mean Infidelity Scores for Different Noise Scale and Resample Probability Combinations for Custom SHAP Method

	Resample Probability		
Noise Scale	0.2	0.5	0.8
0.1	0.077	0.080	0.082
0.5	0.190	0.203	0.209
1.0	0.247	0.258	0.267
2.0	0.299	0.306	0.312

Table 6: Mean Infidelity Scores for Different Noise Scale and Resample Probability Combinations for SVS

	Resample Probability		
Noise Scale	0.2	0.5	0.8
0.1	0.080	0.088	0.094
0.5	0.184	0.193	0.203
1.0	0.249	0.256	0.264
2.0	0.302	0.307	0.311

Table 7: Mean Infidelity Scores for Different Noise Scale and Resample Probability Combinations for DeepLift

	Resample Probability		
Noise Scale	0.2	0.5	0.8
0.1	0.080	0.090	0.097
0.5	0.193	0.201	0.210
1.0	0.257	0.262	0.270
2.0	0.311	0.315	0.319

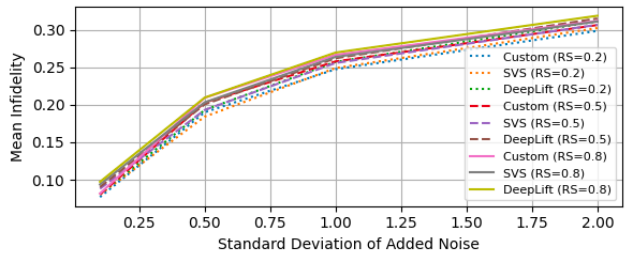

Figure 5: Mean infidelity vs combinations of standard deviation and noise scale, varied by resample probability (RS)