Huazhen Xu

1. Introduction

This report explores two fundamental approaches to model explainability: feature attribution explanations and counterfactual explanations. This report details the implementation and evaluation of both explanation techniques using the Titanic and Dry Bean datasets.

2. Task 1 – Exploratory Data Analysis

2.1. Task 1 (a)(i)

There is a higher survival rate for females comparing to males. This is expected, since females are generally offered priority when boarding life boats. Similarly, lower-class passengers tend to have lower survival rates. On the other hand, there is a strong negative correlation between fare and pclass (=-0.56). This is understandable as better class directly indicates higher ticket fare and class number are inversely assigned according to their qualities. There are also relatively strong correlations between age and pclass. This is expected as older passengers tend to select better classes for the comfort level.

2.2. Task 1 (a)(ii)

Most important features: pclass, sex, fare

Least important feature: sibsp, parch, age

Reason of these choices is based on the common knowledge of the Titanic disaster, as well as correlation analysis between the feature and the dependent variable ‘survived’. For example, Pclass is the feature with strongest correlation with ‘survived’ (-0.3) and sibsp is the feature with weakest correlation (-0.015). However, this analysis only gives information about the data itself, with no direct indications of how particular model decides feature importance.

2.3. Task 1 (a)(iii)

Additional exploratory analysis that can be beneficial for the analysis include:

- Checking feature distributions, identify outliers. Histograms and boxplots can be used to analyse numerical feature such as age and fare. These can help identifying skewed distributions or outlier points (e.g. extremely high fares).
- Using partial dependence plots to visualise the marginal effect of a feature on the predicted survival rate from the model. For example, we can analyse how survival probability changes with changing fare, or how it varies with age.

3. Task 2 – Feature Attribution Explanations

In this section, the various feature attribution explanation approaches are explored and compared for the Titanic dataset and the Dry Bean dataset. The results and analysis are reported below.

3.1. Task 2 (b)(i)

For part (b) of Task 2, I have computed the attribution score with SHAP implementation from previous 2(a), Shapley Value Sampling and DeepLIFT on 10 randomly selected instances from the Titanic test set. The backgound dataset are also randomly selected samples of shape (10,7) from the Titanic test set.

By computing the average of attribution scores per feature for each of the explanation methods, I have observed the following results:

- SHAP implementation:
  - Most important feature: male (1.0844), pclass (0.1074)
  - Least important feature: fare (0.0455), age (0.0575)
- Shapley Value Sampling (SVS):
  - Most important feature: pclass (0.2357), female(0.2136)
  - Least important feature: parch (0.0085), sibsp (0.0486)
- DeepLIFT:
  - Most important feature: fare (0.2288), pclass (0.1122)
  - Least important feature: parch (0.0006), male (0.0349)

Overall, pclass seems to be the most important across all methods, parch seems to be the least important. This aligns with the correlation analysis result from Task 1.

3.2. Task 2 (b)(ii)

The majority of the results aligns across 3 methods. The only substantial difference is on the feature 'male', where SHAP implementation gives the highest score (most important) but DeepLIFT gives the smallest score (least important). The reason for this difference roots from the fundamental differences in the mechanisms of the three methods:

- SHAP distributes importance across correlated features, meaning that male may absorb importance from other related variables.
- DeepLIFT relies on backpropagation and reference baselines, which can cause certain features to receive lower attribution scores if their impact is indirect.
- Shapley Value Sampling introduces randomness, causing variation in attributions across runs.

These differences may introduce discrepancies in importance of each feature assigned.

3.3. Task 2 (b)(iii)

Overall, the attribution scores match my expectation. For most of the cases, pclass and sex received high scores, parch and sibsp received low scores. On some occasions, the attribution scores differ from correlation-based predictions. For example, fare was assigned low importance by SHAP. This could be because SHAP's game-theoretic approach accounts for feature interactions, meaning it does not attribute all importance to a single feature but rather spreads it across correlated variables. Since pclass is already assigned high importance, fare's importance gets partially absorbed by it.

Generally, the reasons for difference between user expected explanation and computed attribution explanation could be that the attributions are computed based on models. If the model did not learn the data distribution correctly, then the attribution score would differ from the expected insights gained by user directly from data.

3.4. Task 2 (b)(iv)

Insights gained from the exploratory data analysis is global and not subject to specific data points. It is fully model-agnostic, and therefore it can provide generalisable insights on arbitrary models trained on the given datasets. However, simple exploratory data analysis may not uncover some model-specific feature importance as it relies on statistical relationships only. It may also neglect complex non-linear relationship between features and target variables.

Feature attribution methods, on the other hand, are specific to particular models and explained data instances. Therefore, they are more likely to capture more unusual model behaviours and more complex relationships, which are usually missing from analysis from data distributions alone. The three methods used each have their own advantages and disadvantages:

- SHAP:
  - Adv: model-agnostic, supports various types of data, provides global and local attributions and can handle complex models well.
  - Disadv: computationally very expensive, and chosen sampling method will affect results and the guarantees.
- SVS:
  - Adv: efficient approximation of Shapley values, easy to interpret.
  - Disadv: introduces randomness, requires many samples for stability
- DeepLIFT:
  - Adv: fast, captures hierarchical interactions in neural networks and interprets activation-based models well.
  - Disadv: sensitive to model structure, not always comparable to SHAP-based methods.

3.5. Task 2 (c)

Choice of infidelity metric hyperparameters:

- **noise_scale** - standard deviation of Gaussian noise for continuous features:
  - **0.05**: low noise, small perturbations, testing the model's sensitivity to minor variations.
  - **0.2**: moderate noise, evaluating the robustness of attributions under stronger input changes.
- **cat_resample_proba** - probability of resampling categorical features:
  - **0.1**: Small chance of resampling, minimal effect on categorical variables
  - **0.5**: Moderate chance of resampling, tests model response to categorical changes

For each combination of parameter values, I have calculated the infidelity score over the entire Titanic test set with the three attribution methods. The resulting infidelity scores are shown below:

	SHAP	SVS	DeepLift
noise=0.05, resample=0.1	0.041132	0.043969	0.038025
noise=0.05, resample=0.5	0.047009	0.05114	0.052279
noise=0.2, resample=0.1	0.115369	0.115901	0.115755
noise=0.2, resample=0.5	0.132789	0.127481	0.133699
Mean Infidelity	0.084075	0.084623	0.08494

We can conclude here that the best method with lowest mean infidelity score is my own SHAP implementation, with score 0.084075, indicating it produces attributions that best align with model behavior across all perturbation settings.

As expected, increasing Gaussian noise (0.05 → 0.2) increases infidelity scores, suggesting the attribution methods become less stable under strong perturbations. Higher categorical resampling probability (0.1 → 0.5) has a smaller effect compared to noise but still increases infidelity scores slightly. DeepLift performs best in one setting (0.05, 0.1) but worsens under categorical resampling, indicating that it may not be as stable when categorical features are perturbed.

3.6. Task 2 (d)

For this part, a separate dataset, Dry Bean dataset, is used to train a separate neural network model. There are 16 input features and 7 classes to be predicted.

Model architecture:

- 2 linear layers of 128 neurons with ReLU activation. Softmax activation for output layer since this is a multi-class classification problem.

- Loss Function: CrossEntropyLoss (multi-class classification)
- Optimizer: AdamW (learning rate = 0.001)
- Training Duration: 1000 epochs

Performance metrics:

- Accuracy: 0.93
- Macro F1 score: 0.94
- Final loss: 1.23
- Per-Class F1 Scores:
    - Seker: 0.9553
    - Barbunya: 1.0000
    - Bombay: 0.9503
    - Cali: 0.9099
    - Dermosan: 0.9554
    - Horoz: 0.9474
    - Sira: 0.8805

Runtime analysis:

for both Titanic and Dry Bean datasets, I have recorded the runtime for each of the three attribution methods on the first 200 samples in the test set. The results are as follows, note the time is in seconds:

	SHAP	SVS	DeepLIFT
Titanic	416.5689	0.1289	0.0125
Dry Bean	> 600 (10 min)	0.3230	0.0178

We can conclude that DeepLIFT is the most computationally efficient method for both Titanic and Dry Bean. SHAP is the least efficient. This is expected as it involves evaluating multiple feature coalitions, leading to exponential complexity. Dry Bean dataset has more features and therefore further increasing the complexity. I therefore have set a time limit of 10 min to stop SHAP execution on Dry Bean dataset if it exceeds the limit. The value of 10 min is chosen since if the execution is not finished after 10 min, we can be certain that SHAP method is much less efficient than the other two.

4. Task 3 – Counterfactual Explanations

4.1. Task 3 (a)(i)

In dataset where different features have different ranges, standard L1 introduces implicit weighting, where features with larger numerical ranges are implicitly treated with higher importance. The normalised L1 distance accounts for this by dividing each feature difference by its range, ensuring that all features are treated equally. In preprocessed dataset, the feature sex is one-hot encoded into two columns (female, male), making it carrying higher implicit weights. In my implementation of distance function, I have balanced this by dividing the weight by two for this feature.

4.2. Task 3 (a)(ii)

To treat each feature equally in the original unprocessed dataset, I have used normalised L1 metric to calculate the distance, ensuring equal feature importance. I have also reduced the weight contribution of each sex-related column by dividing by the number of one-hot columns, ensuring that the sex feature does not get over-weighted relative to other features.

4.3. Task 3 (d)(ii), (e)

For this part, I computed counterfactuals of 20 randomly selected samples from the Titanic test set. Evaluated the mean and standard deviation of validity, proximity and plausibility over 5 trial runs. The results are reported below.

method	validity	cost	plausibility
NNCE	1.0 +- 0.0	0.139 +- 0.018	0.032 +- 0.004
WAC	0.88 +- 0.075	0.194 +- 0.027	0.197 +- 0.027

- Validity Difference:

NNCE validity is 100% (1.0 ± 0.0) because it selects actual training points of the opposite class, ensuring they are always valid counterfactuals.

WAC has lower validity (0.88 ± 0.075) because it relies on a relaxed loss function and gradient-based search, which may fail to find a valid counterfactual in some cases. Possible reasons include local optima, where gradient descent fails to fully optimise the validity term, insufficient iterations, and hyperparameter sensitivity.

- Proximity (Cost) Differences:

NNCE achieves a better (lower) proximity on average (0.139) compared to WAC (0.194). The theoretical expectation is that WAC should have lower proximity, as it optimizes for the nearest valid counterfactual. However, due to local optima struggles and suboptimal hyperparameter tuning, WAC can end up worse than NNCE in practice.

- Plausibility Differences:

NNCE plausibility is significantly better (0.032 vs. 0.197). This is because NNCE counterfactuals are real training points, meaning they naturally follow the distribution of the dataset. WAC counterfactuals are generated using gradient updates, which might not be realistic values, and can produce out-of-distribution samples, which could result in unrealistic counterfactuals. For example, floating numbers in categorical features.