City University London
MSc in Data Science
Project Report
2021

Improving user understanding of Multimodal Models using LIME Explanations

Miguel Bravo

Supervised by: Simone Stumpf

January 2021

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed:  (**Miguel Bravo**)

Table of Contents

Abstract

The aim of this project is to evaluate the effectiveness of AI Explanation techniques – specifically LIME explanations – in helping human users better understand the predictions of multimodal AI models. Using a dataset from Kaggle, a Multimodal AI model is trained to predict pet adoption speed based on online pet adoption listings containing image, text and tabular data. A custom implementation of LIME is then developed to produce multimodal explanations for this model, and a user study is conducted to investigate the effect on user understanding. The main finding is their unexpected negative effect due to significant noise present in the explanations – pointing to the noise present in the model and dataset as the most probable cause. Future work could attempt to address this challenge by selecting a more established and 'solvable' multimodal problem showing greater signal to users. Even so, the current work makes valuable contributions to Explainable AI, representing one of the first known examples of LIME applied to a multimodal problem, and also laying down an effective approach for user evaluation of AI explanations yielding conclusive results.

1 Introduction and Objectives

1.1 Introduction

The purpose of this research project is to evaluate the effectiveness of AI Explanation techniques in helping human users better understand the predictions of multimodal AI models. As part of this, another closely related goal is to develop an effective approach to designing and producing explanations for multimodal models, which are typically not supported by existing techniques. Multimodal models process distinct types of information such as images, text, and tabular data when making predictions. This can make them highly complex and uninterpretable, limiting their applicability for users wanting to take real-life decisions beyond the model's predictions [16].

This project aims to address this challenge through extending and applying a well-established Explanation technique called LIME – Local Interpretable Model-Agnostic Explanations – belonging to the family of post-hoc explanations; techniques enhancing the interpretability of trained models without taking into account their internal mechanics [16]. LIME generates explanations by creating an interpretable 'surrogate' model which approximates the underlying model's behaviour locally around a data point of interest, thereby providing explanations of the underlying model's prediction at this point. This work extends LIME to support multimodal models, since the current implementation can only handle unimodal models.

A between-group user study is then conducted to assess the effect of the LIME explanations on participants' understanding of how a multimodal model makes predictions. Statistical analysis will be conducted on the data collected in the study to infer whether LIME has a statistically significant positive effect on user understanding. There will then be a critical evaluation of the main results obtained to draw out the key findings and provide a direction for future work.

1.2 Problem Background

The specific problem explored in this project is a multimodal classifier predicting the adoption speed of pets on the website PetFinder.my, based on their online listing information which includes images, text and tabular data. This problem and the dataset were sourced from a public Kaggle competition [11].

This problem was chosen due to its multimodal nature, in line with the project's research aims, and also its value from a real-world applicability standpoint. It is an example of the more general task of helping users better understand and therefore optimise an online listing in order to drive a specific outcome of interest. This has wider applications beyond online pet adoption, including online property rentals, dating platforms, and job recruitment websites – to name a few examples.

In addition, another key reason is that the problem was sourced from a Kaggle competition. This provided reassurances around the quality and tractability of the problem and the data since it had already been tackled by thousands of AI practitioners with their results and solutions published online. It also allowed for the modelling in this project to build upon an existing knowledge base, and provided a natural performance benchmark based on the competition's leading submissions.

Finally, the size, quality, and variety of the dataset – consisting of +14k examples of pet listing images, text, and tabular data – made it quite a rare find, which is another key reason for this choice of problem. This allowed for a broad range of techniques to be considered and applied in the modelling phase of the project. Section 1.2.1 explains the specifics of this dataset in more detail.

1.2.1 Pet Adoption Dataset

The training dataset for this problem – downloadable from the Kaggle competition webpage [12] – consists of over 14k historic pet adoption listings on the website PetFinder.my, each labelled with the pet's eventual speed of adoption since the listing was posted, which could be any of the following categories: *Same day* adoption; *0-7 days*; *8-30 days*; *31-90 days*; and *100+ days*. Pet listings include images of the pet, a text description, and tabular data on the pet's key attributes (e.g. age, breed and so on), as illustrated in Figure 1 below. The prediction task therefore requires training a multimodal classifier which takes the different data sources in a pet's listing as input, and predicts the pet's adoption speed category based on this information.

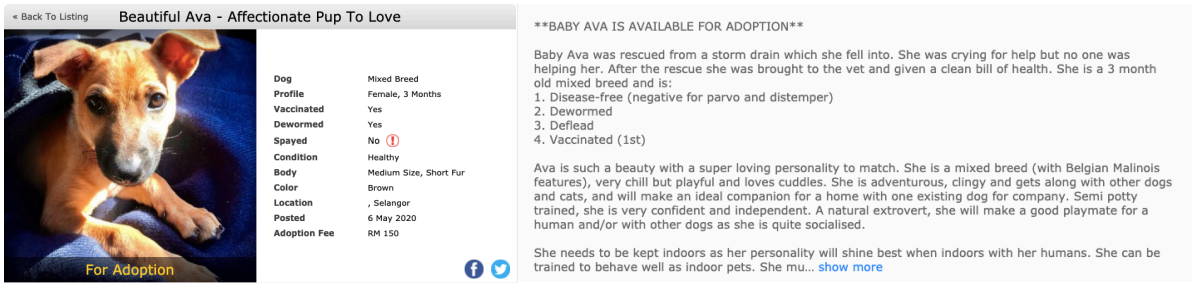


Figure 1: Example pet adoption listing on PetFinder.my

1.3 Objectives

Based on the foregoing discussion, the main research question can be expressed as follows:

> **How can we develop explanations for multi-modal AI models to enhance user understanding?**

Table 1 shows the key research objectives defined for this project in order to answer this question, along with the relevant tests that show whether each objective has been met.

#	Project Objective	Test
1	**Modelling**: Pre-processing training data and developing a multimodal model achieving an acceptable performance threshold	The model's performance ranks within the top 100 submissions for the Kaggle Competition from which the problem and dataset were sourced.
2	**Explainability:** Applying LIME explanations for each input type in the model (e.g. image, text, tabular), in a way that is interpretable to users.	Visual explanations are successfully produced for each data type, which are easy to understand for users
3	**User Evaluation:** Conducting a user study to understand the effect of LIME explanations on users' understanding of the model's predictions, in order to answer the research question.	Statistical Hypothesis Testing to determine whether LIME has a positive effect on user understanding, which is statistically significant.

Table 1: Key objectives to answer the main research question, and test conditions to determine whether these have been met

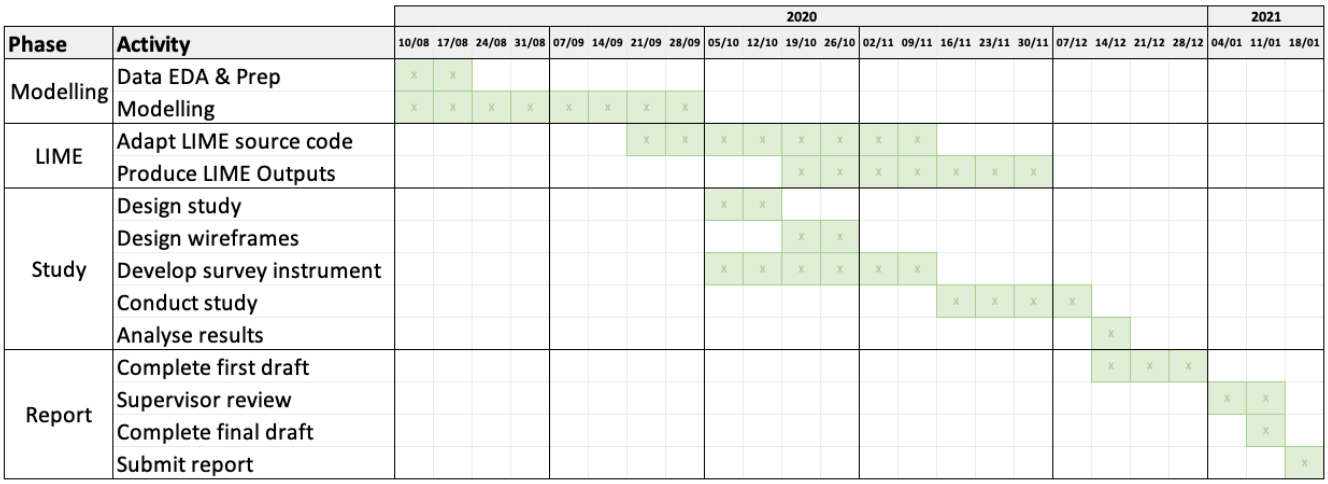
1.4 Project Beneficiaries

This work can be seen to benefit a wide range of groups both technical and non-technical, with either a research or practical interest in making models more interpretable. Below is an overview of the main beneficiaries identified for this project:

- **Non-technical** users wanting to understand a model's predictions to inform and improve their own decision making to achieve a desired practical outcome [16]. For instance, in the context of this problem, such users may ask themselves: What are the key factors affecting a pet's adoption speed? How can I alter the listing to improve adoption speed?
- **Researchers** interested in applying LIME explanations to multimodal models, as this is not supported in the current implementations of LIME, and was developed specifically for this work – representing one of the key contributions of this project.
- **Designers** looking to explore how to digitally present LIME explanations to users and the main pitfalls to avoid – based on the design work conducted as part of the user study component of the project. This can help inform their efforts in designing AI explainability interfaces for more effective human-computer interaction.

1.5 Work Plan

Below is the workplan devised for this project, organised by project phase – broadly aligning with the overall objectives – and its constituent activities, along with relevant timescales. It provides a snapshot at-a-glance overview of how the project was conducted end-to-end, how long it took, and how the effort was apportioned across the different tasks.



1.6 Overview of Report Structure

The report has been structured to clearly highlight and explain each stage of the research process undertaken, from conducting background research detailed in the Context section, to performing the analysis and interpreting the results – as set out in the Methods and Results section – to critically evaluating the key findings and lessons learned of the project, explained in the Discussion & Evaluation section. Below is an overview of this structure, and the purpose of each section:

- **Context:** Survey of the literature relating to the key research areas this project pertains to: Multimodal AI; AI Explainability; and User Evaluation. Foundational theory and concepts are raised here, along with research gaps this work aims to address.
- **Methods:** In-depth look at the methods used to carry out this project, from producing a multimodal AI model, to applying LIME explainers, and conducting a user study. Sufficient detail is provided so that the reader can replicate the project end-to-end.
- **Results:** Summary of the main outputs and results obtained from each of the methods, as they relate to each of the key project phases; Modelling, Explainability, and Human Evaluation. Emphasis is given on the most interesting and insightful findings, with figures included to help explain results as appropriate.
- **Discussion & Evaluation:** Exploration of the results, interpreting their significance within the context of the project objectives in order to draw out the main conclusions from this work. This section also includes an evaluation of the project as a whole, to identify lessons learned and inform future work.

2 Context

This project involves applying explainability techniques to help interpret an AI model taking different data sources as input, and then evaluating these techniques through a user study. It can therefore be said to lie at the intersection of three different research areas – XAI, AI, and HCI (Human-Computer Interaction).

This section sets out the current research landscape across these three fields as they relate to the main goals of the project, discussing how the project fits into previous work conducted in these areas, while also highlighting the research gaps being addressed. In particular, little research seems to have been done on applying model-agnostic explainability techniques such as LIME and SHAP to multimodal models – representing one of the key contributions of this project.

2.1 Multimodal AI

Multimodal AI involves building models that can process and relate information from multiple sources of data or modalities. These are different ways that data can be captured on the same event, whether that is through natural language (e.g. text), visual signals (e.g. images), audio signals (e.g. recordings), or scalar measurements (e.g. tabular data). A multimodal model needs to be able to make sense of and combine these different inputs when making predictions [2].

This field holds great potential since the world around us is multimodal, and models that exploit this can offer richer insights into complex real-world tasks. However, this comes with significant challenges given the heterogeneity of multimodal data. The main challenge, relevant to this project, is how to represent and summarise multimodal data in a format that a computational model can work with, exploiting the complementarity of each data source and their varying predictive power to enhance the model's performance [2]. In what follows, this section will delve in more detail into the two main types of multimodal representations – *joint* and *coordinated* representations.

2.1.1 Joint Representations

Joint representations relate to any approach which aims to combine and project the different sources of data into the same representation space. In recent years, the most common

example is a concatenation of individual modality features, using neural networks (NN). Due to the multilayer nature of NNs, each successive layer represents data in a more abstract way, and the final or penultimate layer can then be used as a data representation.

Below are the main ways NNs can be used to produce joint representations for prediction tasks, which will be considered in the modelling phase of the project:

- **Early Fusion:** NNs produce separate data representations for each modality which are then concatenated together immediately after extraction to form a joint representation. This is the simplest method to implement, but fails to account for the varying informativeness of each modality in how they are combined, instead giving them equal weighting [5].
- **Mid-level fusion:** A single NN extracts representations from each modality using separate hidden layer structures, followed by a set of shared hidden layers to fuse these representations into a single learned representation [1]. While more complex to implement, this has the advantage of producing a single learned representation capturing the varying informativeness of each modality
- **End-to-End:** A single NN is used end-to-end not only to extract a joint representation – as with mid-level fusion – but also to perform the further multimodal prediction task in question [2]. Figure 2 shows an example of this, with a deep neural network model performing an image-to-text translation task [3]. The model implements initial feature extraction, mid-level fusion, and prediction end-to-end.

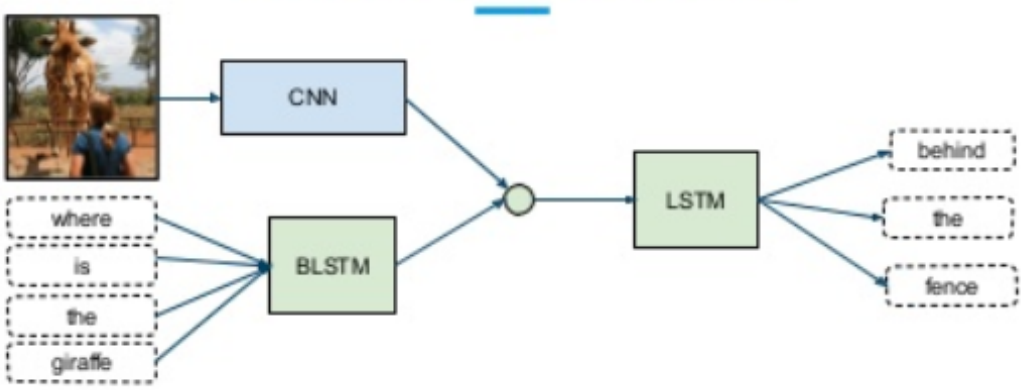


Figure 2: Example of a multimodal deep learning model performing an image-to-text translation task end-to-end, from feature extraction through to prediction.

2.1.2 Coordinated Representations

Coordinated representations involve learning separate representations for each modality coordinated through a common constraint – rather than attempting to project modalities to a single space like joint representations. Broadly speaking, there are two main approaches when producing such representations:

- **Similarity across Spaces:** Creating representations that minimize the distance between similar concepts across the coordinated spaces. For example, such models would create a representation of the word ‘dog’ which will be closer to a representation of the image of a dog compared to the image of a car [10].
- **Structured Coordinated Spaces:** Creating representations that impose additional constraints on the coordinated representations beyond similarity. Different constraints are applied depending on the task in question, such as hashing, cross-modal retrieval and image captioning [5].

Again, NNs have proven highly effective for producing coordinated representations, due to their ability to learn representations directly from raw data, which are more informative than hand-crafted representations.

For both joint and co-ordinated representation learning, the main disadvantage of using neural network approaches is their lack of interpretability. However, this should not pose a problem for the purposes of this project, since separate explanation techniques will be applied as part of the analysis. On the contrary, this highlights the need for the work carried out in this project to make multimodal representations and models more interpretable.

One final challenge to note is that neural networks require large training datasets to be successful [2]. There is a risk that neural network based representations may not exceed other more traditional feature extraction processes on medium-sized datasets, such as the one used in this project. To address this, one solution is to use pre-trained neural network models trained on massive datasets ahead of time, which can be repurposed for the task at hand. This approach will be explored in this project.

2.2 AI Explainability

Explainable AI refers to the application of methods and techniques to make AI systems understandable to human users. This field has risen in prominence in recent years as

breakthroughs in AI systems have increased their predictive power, at the expense of massively increasing their complexity to the point where they come to be regarded as uninterpretable black-box decision engines – underscoring the need for research to make such systems more transparent [9].

This has sparked renewed research interest into making AI systems more understandable, especially as they are increasingly used for real-world tasks requiring them to satisfy additional 'auxiliary' criteria beyond simply prediction accuracy, which ultimately require a better understanding of how they make decisions. Some key examples are summarised below:

- **Fairness:** With AI models increasingly used to inform decisions impacting human lives – from credit applications, to court sentencing – they need to be monitored for fairness to ensure they do not unfairly discriminate against individuals. AI interpretability is required here to identify and correct such issues when they arise [8].
- **Causality:** AI models can be used to uncover potential causal relationships – based on the associations they learn from the data – to inform new hypotheses for researchers to test experimentally – thereby serving as a valuable knowledge discovery tool. AI interpretability is required here for practitioners to be able to identify the associations learned by models [16].
- **Informativeness:** AI models can also be used to provide extra information to practitioners in addition to predictions, which helps them to better understand a specific domain and make better human decisions as a result. Again, this requires interpretability in order to extract additional useful information about how the models work [16].

Informativeness is the main auxiliary criteria addressed in this project, given the research goal of applying AI explanations aimed at improving user understanding and informing them on the key drivers influencing a model's predictions.

2.2.1 AI Explainability – Taxonomy

AI Explanations fall into two main types – intrinsic and post-hoc. Intrinsic interpretability refers to ML models that are considered inherently interpretable due to their simple structure, such as short decision trees or sparse linear models [21]. Simply by studying their internal logic, users can understand how these models work to support auxiliary real-world tasks. For

instance, in a sparse linear regression model, the learned parameters represent the impact of each of the features on the models' predictions, making the model inherently interpretable.

Post-hoc interpretability refers to the application of separate explanation methods after models have been trained [21], to make them understandable without needing to elucidate their internal logic. These techniques are suited for highly complex models with high predictive power – a great option to satisfy auxiliary tasks without sacrificing on a model's performance, such as deep learning AI models.

The model to be considered in this project will be multimodal and require high complexity and predictive power to successfully learn from the different data sources. Therefore, post-hoc explanation methods seem most appropriate for this research project. In what follows, some of the most established post-hoc techniques are summarised in more detail, along with their pros and cons, as they relate to the project.

2.2.2 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a model-agnostic explanation technique developed by Ribeiro et al. [23], involving the use of local surrogate models to explain individual predictions of black box AI models. These models are trained to locally approximate individual predictions of the underlying black box model, and provide explanations by being inherently interpretable models themselves.

LIME generates a local surrogate model by first creating a permuted dataset, consisting of permuted samples and the corresponding predictions of the black-box model, taken from a defined neighbourhood around the instance of interest. An interpretable model is then trained on this dataset, with points weighted by their proximity to the instance of interest. Any model can be used so long as it is interpretable – such as a Decision Tree or Linear model – in order to provide explanations of the prediction. The main requirement for such models is local fidelity – i.e. they should be a good approximation of the underlying model's predictions locally, so that their explanations are accurate [23].

Figure 3 illustrates the process of applying LIME to a model processing tabular data, showing how points are sampled and used to train a linear surrogate model approximating the underlying model's predictions locally around the instance of interest [23].

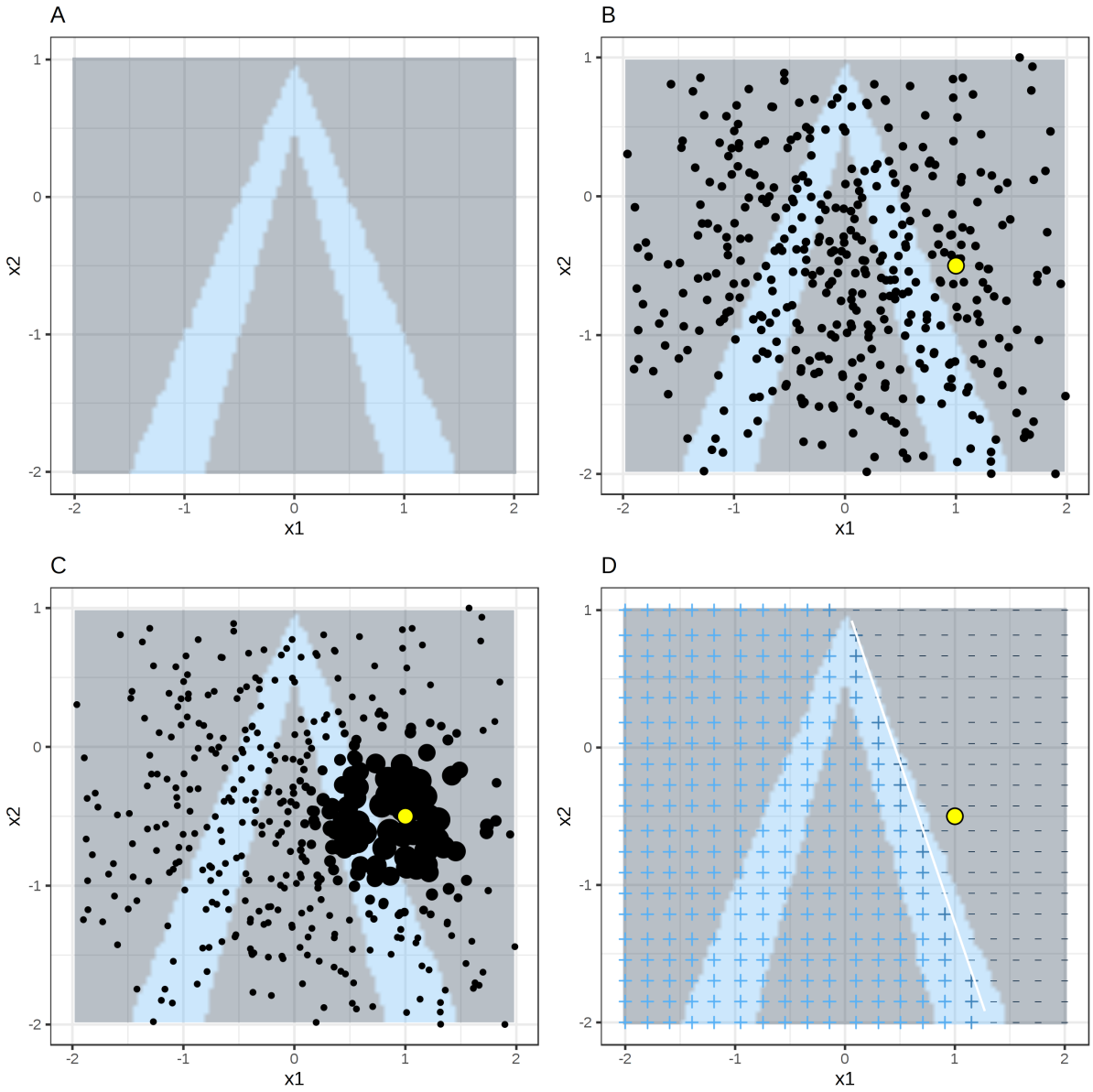


Figure 3: A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark) or 0 (light). B) Instance of interest (big dot) and data sampled from a normal distribution (small dots). C) Assign higher weight to points near the instance of interest, D) Signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

In addition to tabular data, LIME can also be applied to images and text, making it highly suitable for the multimodal problem being addressed in this project. With text, the permuted dataset is generated by randomly adding or removing words from the original text, represented as binary features indicating whether each word is present. For images, the image is first segmented into 'superpixels' – groups of pixels that share common characteristics – and the permuted dataset is generated by turning these superpixels on or off. These are switched off by replacing them with a uniform colour such as gray [21].

Other benefits of LIME include its model-agnostic nature. LIME can be applied to any underlying black box model using the same local surrogate model, meaning the same

corresponding explanations can be generated for any model no matter its internal complexity. Also, with simple enough local surrogate models, the resulting explanations become more human-friendly – shorter and easier to understand especially for non-technical users [21].

The main drawback is the instability of LIME explanations. For a given instance, running LIME multiple times can result in different explanations, as a different permuted dataset will be sampled each time which can affect what the local surrogate model learns. In addition, for tabular data, the permuted dataset is heavily dependent on the kernel size of the sampling neighbourhood and it is not clear how to set this optimally, other than through manual trial and error. This instability may undermine the fidelity of these explanations, and so should be treated with caution [21].

2.2.3 Shapley Additive Explanations (SHAP)

SHAP is a method developed by Lundberg and Lee [18] to explain individual predictions of a model by treating the contribution of each feature value to the prediction as Shapley values from Coalitional Game Theory. In SHAP, Shapley values are estimated as the learned feature weights of a local surrogate model trained on a permuted sample around the point of interest (similar to LIME). The specific method followed to create the sample – KernelSHAP or the more compute-efficient TreeSHAP [19] for tree-based models – is what ensures that the learned weights of the surrogate model are Shapley values, which provides a number of key benefits:

- Thanks to the properties of Shapley values – derived from Game Theory – SHAP's learned feature contributions are always guaranteed to be distributed fairly, in that they always add up to the overall difference of the individual prediction relative to the model's average prediction. This makes SHAP explanations internally consistent and stable, unlike LIME.
- Thanks to this local consistency, SHAP is able to aggregate up local Shapley values to generate sensible global explanations for a model's behaviour; which again is not possible with LIME. This makes SHAP well suited for providing non-technical users with a high-level overview of a model's key prediction drivers overall.

That said, SHAP also comes with certain limitations compared to LIME. First, SHAP is slower and more computationally intensive, as Shapley values are expensive to calculate even with the estimation methods devised for SHAP. Second, the need to estimate Shapley feature

weights places constraints on how data points are sampled to train the surrogate models, which can result in unlikely points being used and unintuitive weights learned. In addition, these constraints also make the open-source SHAP implementation more difficult to adapt to handle multimodal models, as required by this project. Finally, it is not as straightforward to apply SHAP to different data types compared to LIME, especially text data, which may be problematic for this project [21].

2.2.4 Counterfactual Explanations

A counterfactual explanation is a model-agnostic technique used to explain predictions of individual instances by describing the smallest change to the feature values of a model that changes its prediction to a predefined output of interest. This has the advantage of producing human-friendly explanations that are informative and practical in that they tell users exactly which changes to make to the task being modelled, to get the desired outcome of interest (as predicted by the model). For instance, in a loan application context, this explanation can tell the applicant what they would need to change in their profile for their loan to be approved [21].

Counterfactuals have been shown to work well with tabular data [26]. However, they are challenging to apply on image and text – both within scope in this project – given their high dimensionality making it difficult to generate plausible counterfactual explanations. Martens et al. have come up with an approach for text involving finding a minimal set of words such that removing all words from this set changes the predicted class [20]. Liu et al. have devised an approach for images using a GAN to generate plausible reconstructions by altering image style parameters [17]. However these approaches are difficult to implement, with no out-of-the-box implementation available – making Counterfactuals the least feasible explainer to apply within this research project.

2.2.5 Multimodal AI Explainability

AI explainability in a multimodal setting has mostly focused on deep learning models that produce visual and textual explanations alongside their predictions. One common use case for these models relates to visual question answering prediction tasks. For instance the work carried out by Park et al. [22] involved a deep learning model capable of answering questions about images by pointing to the relevant parts of the image influencing its answer, and an accompanying textual justification – as illustrated in Figure 4 below.

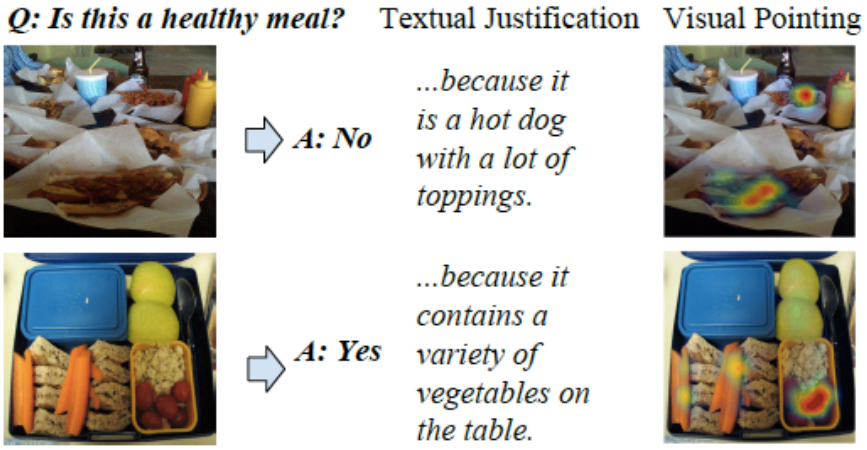


Figure 4: For a given question and an image, the Pointing and Justification Explanation (PJ-X) model developed by Park et al. predicts the answer and multimodal explanations which both point to the visual evidence for a decision and provide textual justifications.

Another setting where this type of technique has been applied is in video classification. Recent work done by Kanehira et al. [13] consisted of creating a video classification deep learning model capable not only of predicting a video's class (i.e. what it's about) but also generating visual and textual counterfactual explanations of why it was not classified to a different class, as shown in Figure 5.

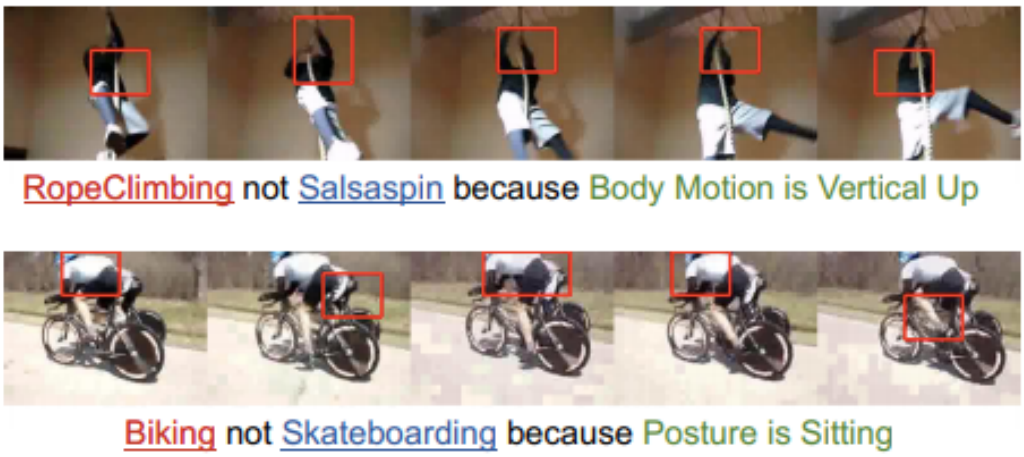


Figure 5: Example output from the video classification model by Kanehira et al. providing visual and textual explanations of why the model favoured one class over another in its prediction.

While this research shows great promise in creating multimodal explanations, it differs from this project in one important respect. The explanation techniques used are built into the architecture of the prediction models being explained, and so cannot be easily applied to other AI models; i.e. they are not model-agnostic. By contrast, the explanation techniques explored in this work, such as LIME and SHAP, are model-agnostic in that they can be

applied to any model regardless of its internal architecture. In this respect, this project hopes to make a novel contribution to the area of multimodal AI explainability.

2.3 Human Evaluation

Human evaluation of ML explanations is essential to assess their interpretability for a specific task. Therefore this research project will necessarily include a user study with human subjects to test the quality of the Counterfactual and SHAP explanations. According to Doshi-Velez and Kim [9], there are two different types of human evaluation for ML explanations; application-grounded and human-grounded. Application-grounded evaluation involves conducting human experiments with a real application and real expert users, to test whether the system delivers on its intended task. Human-grounded evaluation is about conducting simpler experiments with lay users – not necessarily real users – to test the general quality of explanations, allowing for cheaper and more feasible experiments with access to a larger subject pool [9]. For these reasons, a human-grounded evaluation approach seems most appropriate in this research project, given the limited time and resources available.

Another key question is the kind of interface to present users with AI explanations as part of the user evaluation. One effective approach by Kulesza et al. [15] involves producing a software application showing interactive explanations of an AI system, allowing users to get a better understanding of how the system made predictions through active exploration. Figure 6 gives an overview of this software. Users exposed to the software's interactivity and graphical outputs achieved 50% better understanding of the AI system, compared to the control group – highlighting the effectiveness of an interactive interface. However, such functionality may prove overly challenging to implement for this project.

Alternatively, a separate empirical study by Dodge et al. [8] – aimed at exploring how AI explanations affect users' perception of a model's fairness – used simpler explanation outputs, consisting of static text-based explanations, as illustrated in Figure 7. While it did not allow for the same level of user engagement, the simplicity of the outputs makes it a more feasible interface to implement. The explanation outputs in this project will seek to find a middle ground between both these approaches, balancing effectiveness with ease of implementation.

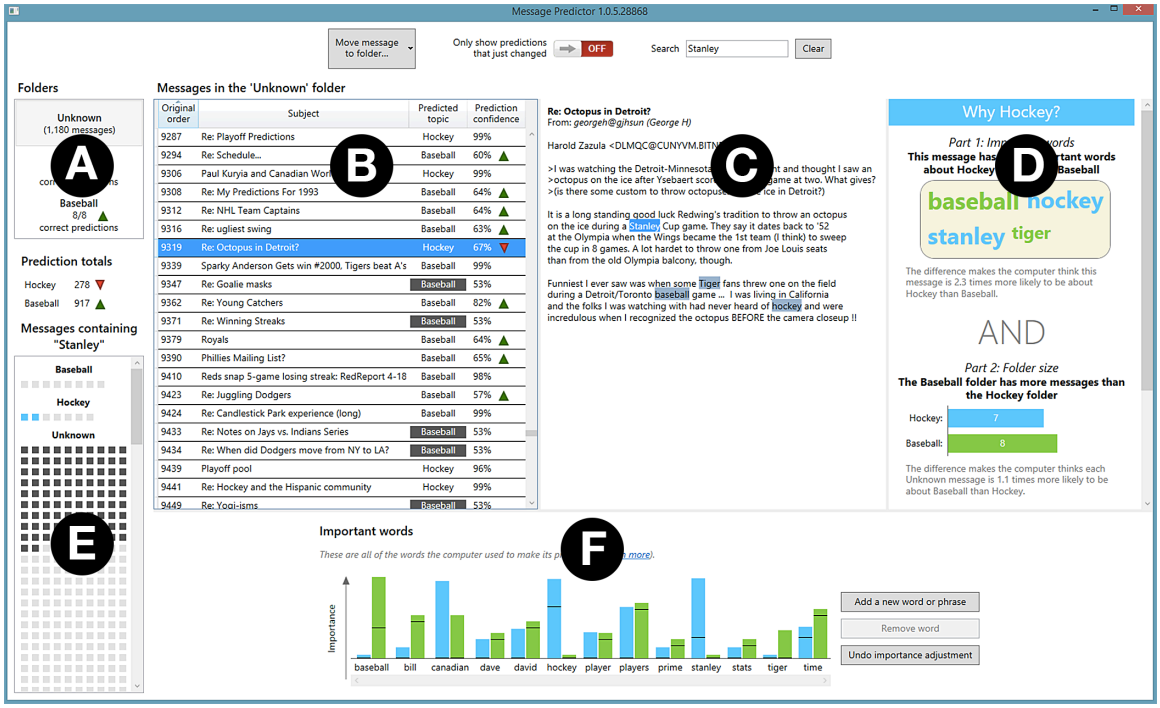


Figure 6: The EluciDebug software. (A) List of folders. (B) List of messages in the selected folder. (C) The selected message. (D) Explanation of the selected message's predicted folder. (E) Overview of which messages contain the selected word. (F) Complete list of words the learning system uses to make predictions.

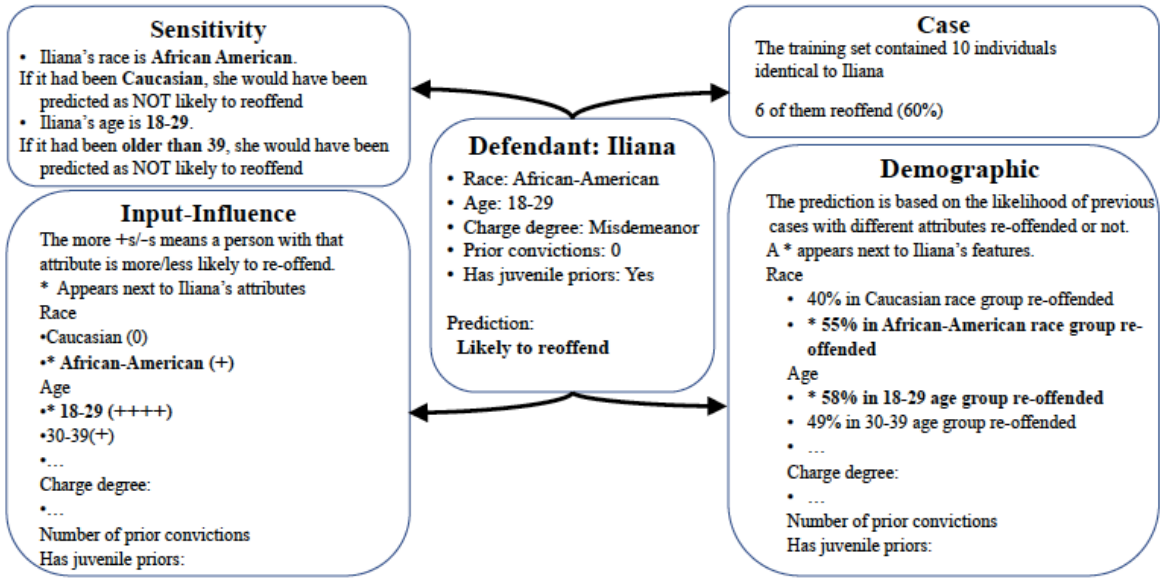


Figure 7: Examples of different types of static AI explanation outputs generated as part of the empirical study conducted by Dodge et al. into the impact of explanations on users' fairness judgements of a model's predictions.

3 Methods

The methods used over the course of the research project align to the key activity areas needed to complete each project objective (as set out in 1.3) and ultimately answer the main research question (as set out in 1.1). The high-level activity areas and the main methods used to complete them are summarised below:

- Pre-processing data and training a multimodal AI model
- Applying LIME Explanations to explain the model's predictions
- Conducting a Study to understand the effect of LIME explanations on users' understanding of the model's predictions
- Analysing the Study data to confirm whether LIME explanations have a positive effect on users' understanding which is statistically significant

The remainder of this section delves into each of these activity areas in more detail, explaining the specific methods used, how they were applied, and the rationale for selecting them from other available alternatives. The aim is to provide enough practical detail to be able to replicate the delivery of the project from start to finish.

3.1 Producing the AI Model

3.1.1 Feature Extraction

The first step was to extract features from the pet listing information – in preparation for training the model – which the model would find useful in learning how to predict adoption speed. As explained in 2.1, an early fusion approach was taken to generate joint multimodal representations as input to the model. Each modality was pre-processed separately, followed by a simple concatenation of the extracted features from each source to create the final representation.

It is worth noting this approach did not account for the varying predictiveness of each modality. There are alternative techniques taking account of this, such as training a single deep learning model to extract features from each source, and produce an optimal weighted combination of these features, where the weights are learned parameters of the model [3]. However, unweighted concatenation was used instead due to its more straightforward

implementation and because it was the preferred method of most submissions in the Kaggle competition [24], lending credibility to this simpler approach.

The specific feature extraction and other pre-processing techniques applied to each data source – and the methods and tools used – are explained in more detail below.

Image Representation

Each pet listing in the training dataset could contain multiple pet images in different sizes, or else no image at all. To simplify the processing pipeline, only the main profile image was taken into account with the remainder discarded, and a black image was used instead where a listing had no image available.

Features were then extracted using a pre-trained DenseNet121 neural network – downloaded from the torchvision python library – trained on the ImageNet dataset. This model was used due to its ability to extract rich image representations for image classification tasks. Its final classification layer was removed so that it output 1024-dimensional image representations. The pet listing images were then resized to 224x224 pixels and normalised, before being passed through the model to extract these representations. Finally, 1D average pooling – with kernel size 4 – was applied to reduce these to 256-dimensional feature vectors.

This process was implemented in python using the pytorch deep learning framework. The compute time proved intractable on a CPU machine, due to the complexity of the DenseNet model and the large number of images to process. The code was therefore executed on Google Colaboratory to make use of the platform's free GPU resources, speeding up compute time significantly.

Text Representation

The pet text descriptions varied considerably from one listing to another in terms of length, content, and the use of punctuation and special characters. The first processing step was to 'clean up' the descriptions, removing among other things: capitalisation; punctuation marks; web URLs; and carriage returns. This was to make the texts more standardised to improve the results of feature extraction.

Feature extraction was then performed using a pre-trained neural network model called Sentence-BERT, trained on the SNLI [4] and the multi-genre NLI datasets [27] – over 1m

sentence pairs labelled with textual entailment information. This allows the model to derive semantically meaningful sentence representations, where similar sentences are close in vector space – making them more semantically separable and distinguishable for AI models. The pet text descriptions were passed through this model to extract 768-dimensional feature representations. A TruncatedSVD algorithm was then applied to reduce these to 67-dimensional feature vectors, preserving 90% of the original vector variance.

The Sentence-BERT model was downloaded from the sentence-transformers python library. It was implemented using the pytorch framework, on Google Colab to make use of the GPU resource, to accelerate compute time.

Tabular Representation

The tabular data consisted of different pet attributes such as Age, Fur Length, Gender, Breed and so on. These required minimal pre-processing as they are already in a useable format by AI models as tabular data. The only step was to standardise the numerical features so they would have equal weighting on the model. The categorical features were used directly without any pre-processing (e.g. one-hot encoding), as the algorithm used for modelling, LightGBM, – explained in more detail in 3.1.2 – has in-built handling of categorical features.

Figure 8 shows a diagram visually summarising the foregoing discussion, illustrating the end-to-end pre-processing pipeline to produce the final feature set used as input to the model.

3.1.2 Modelling

The AI model used was LightGBM, belonging to the family of Boosted Decision Tree ensemble algorithms. This method works by combining many ‘weak’ learners (i.e. trees) into a strong model by adding a tree iteratively at each step that improves the performance of the overall ensemble. LightGBM is a highly optimised type of Boosted Decision Tree, with features making it more efficient to train and improving performance, such as Histogram-based splitting of trees, ignoring sparse inputs, and subsampling inputs [14]. This appears to be the most popular algorithm used in the Kaggle competition, which is the main reason it was employed for this research project.

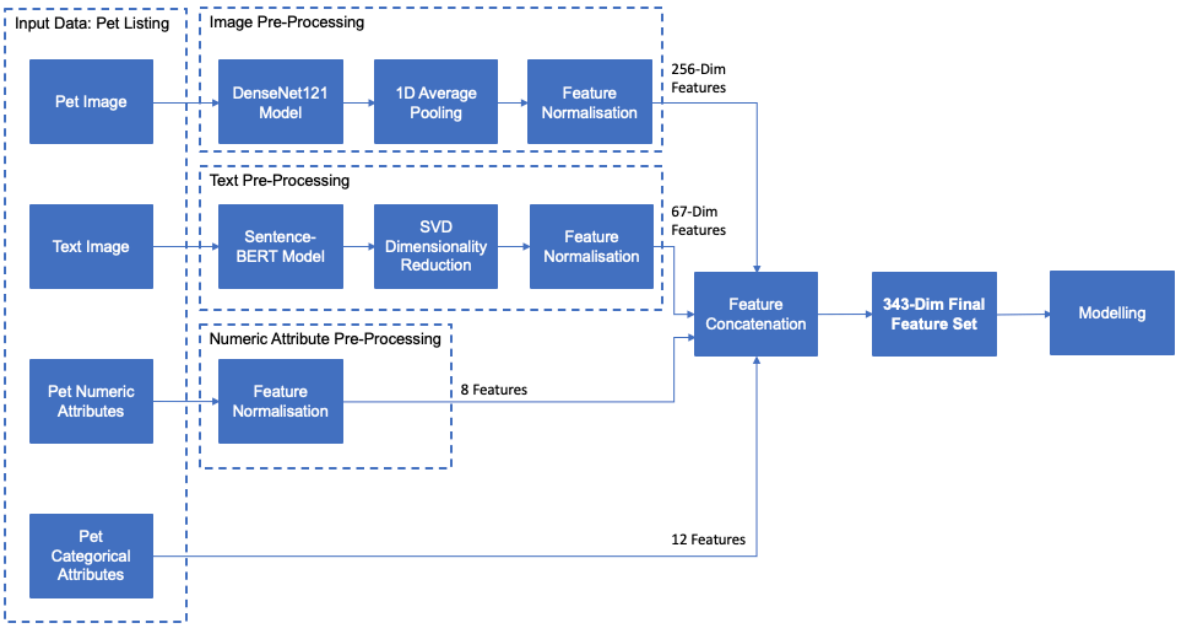


Figure 8: Diagram showing end-to-end pre-processing pipeline illustrating how features are extracted separately for each modality and concatenated together to form the joint representation for downstream modelling.

An alternative approach initially considered was using a multimodal deep learning network capable of performing end-to-end feature extraction from the different modalities (i.e. data sources) followed by prediction. One of the key benefits is its ability to learn better feature representations from each source, by having multiple sources present at feature learning time. However this method was ultimately discarded as it was more complex to implement, and because it was unlikely to outperform LightGBM given the relatively standard size of the training dataset available – typically deep learning models require much larger datasets to significantly outperform other methods.

To simplify the modelling task, the number of prediction speed target classes was reduced to 4, by combining the *Same Day* adoption, and *1-7 Days* categories into a new category called *0-7 Days.* The final class labels used for modelling were therefore: *0-7 Days, 8-30 Days, 31-90 Days, +100 Days.* The intention here was to decrease the complexity of the prediction task, thereby increasing the performance of the model.

There was some class imbalance in the training dataset, with a noticeable skew towards pet adoption speed in *8-30 Days* or *+100 Days,* which risked making the AI over-predict these class labels. To avoid this, the classes were rebalanced by over-sampling the under-represented classes – i.e. adding copies of instances with replacement.

Each of these labels were then encoded with numerical values 0, 1, 2, 3, in ascending order of pet adoption speed. The prediction problem, though a classification task as explained in 1.2, was then effectively treated as a regression problem, training the LightGBM algorithm to predict a scalar value, which is then rounded to one of the numerical class values, using optimal rounding thresholds. These thresholds were optimised to maximise the model's validation performance, using the Nelder-Mead optimisation algorithm.

3.1.3 Model Evaluation

The evaluation metric used to assess model performance was the Quadratic Weighted Kappa (QWK), which is the metric used in the Kaggle competition, allowing benchmarking of the classifier against the competition's top submissions. The Kappa coefficient is a metric which measures agreement on nominal scales between two raters – in this scenario, between the model's predictions and the ground truth. As the name suggests, the Quadratic Weighted Kappa (QWK) is a variant which applies a weighted penalty on disagreements growing quadratically in line with the difference between the prediction and ground truth [25]. It is therefore the appropriate metric for this problem, as it accounts for different degrees of misclassification when the model predicts adoption speed. For instance, if a pet's true adoption speed is 0-7 days, predicting 31-90 days is a larger error than predicting 8-30 days, which is reflected in the QWK score.

The model was trained using a 80%/20% train/test set split of the data consisting of 11,994 and 2,999 instances respectively. During training, 5% of the training set was held out for validation purposes to evaluate model performance over different hyperparameter selections and select the best performing model configuration. Once the best hyperparameters had been found, the final model was then evaluated on the test set to estimate the final generalisation performance (i.e. QWK score).

3.2 Applying LIME Explanations

3.2.1 LIME Implementation

After setting up the model, the next stage was to apply the LIME algorithm to explain the model's predictions based on the input data. This represents one of the project's key original contributions, in that no other previous work seems to have focused on applying LIME – or other model-agnostic post-hoc explainers – to explaining multimodal models. Indeed, the

open-source packages for implementing the well-known explainers LIME and SHAP, only support single-input models focusing on either image, text or tabular data but not any combination thereof. This is surprising given that multimodal models tend to be highly complex, and therefore natural candidates for post-hoc explanation methods.

Apart from LIME, other well-established post-hoc explainers were initially considered for this project including SHAP and Counterfactuals. However, these were eventually dismissed in favour of LIME due to implementation challenges they faced for this task. Like LIME, SHAP would have required source code changes to its implementation library to support multimodal models, which would've been significantly more challenging given its greater algorithmic complexity as explained in 2.2.3. With Counterfactuals, no implementation currently exists which adequately supports high-dimensional image and text data – such as the kind used in this project – and so was discounted.

The main challenge in applying LIME was to adapt the original lime package to support the multimodal model for this project. This meant adapting the different unimodal methods for producing explanations – the image, text and tabular explainers – so that they can take all modalities present in each pet listing as input rather than just the single modality for which they were designed. Each explainer was repurposed to take permuted samples from its relevant modality while holding the other modalities fixed at the instance of interest. This gave a full permuted dataset with all features present to generate the corresponding predictions of the underlying model, and thereby train the local surrogate model to generate explanations for the relevant modality.

The changes were implemented by creating a custom package inheriting from the original explanation methods and then applying the necessary updates to the source code. The relevant code is included in Appendix B.

3.3 Conducting User Study

With the model and LIME explanations in place, all the key AI components were now available to conduct the User Study to answer the main research question – whether LIME explanations improve users' understanding of how the multimodal model makes predictions. The main steps required to carry this out were:

- **Study Design:** Designing the study procedure to collect and analyse the data to

answer the research question.

- **Study Development:** Designing and developing the instrument for users to participate in the study
- **Conducting Study:** Recruiting participants, conducting the study and collecting data for analysis.

The following sub-sections explain the methods used in each of these steps in greater detail.

3.3.1 Study Design

The Study design consisted of a between-group, single-factor experimental set-up, with the factor varied being an experiment condition: whether or not to provide LIME explanations to help users understand the model's predictions. The treatment group would have access to these explanations, while the control group would not. In this way, the study would then enable statistical analysis to infer whether LIME explanations improve user understanding, by comparing data collected from the treatment and control groups using statistical hypothesis testing.

The concept of 'user understanding' was defined as the user's ability to simulate the predictions of a model based only on the input data available to the model at prediction time. This would be measured by calculating the accuracy of the user versus model predictions, using the Quadratic Weighted Kappa score as the metric – for reasons explained in 3.1.3.

70 participants were recruited for the study, and evenly assigned at random into either the control or treatment group for the study. This sample size was based on the expected study parameters, in order to achieve the required 0.8 level of power, as shown in Figure 9.

Parameter	**Expected Value**
Treatment Group – Mean Weighted Kappa	0.7 +/- 0.3
Control Group – Mean Weighted Kappa	0.5
Alpha	0.05
Required Power	0.8

Figure 9: Expected parameters of the Study. This is used to calculate the sample size required to conduct the Study

The above design considerations were then used to inform the following procedure for participants to follow as part of the study, to collect the data required to measure their understanding and answer the research question:

1. **Group Allocation:** Participant is randomly assigned to either the treatment or control group, with 35 participants per group.
2. **Training Phase:** Participant is shown 12 listing examples and the model's relevant prediction (3 examples per predicted class), to familiarise them with how the model makes predictions. If they are in the treatment group, they are shown both pet listing details and LIME explanations for each example. If assigned to the control group, they are only shown pet listing details.
3. **Prediction Phase:** Participant is then shown 10 new listing examples and asked to use their understanding developed in the Training Phase to estimate the model's prediction for each example. These collected datapoints would then be used to calculate the participants' QWK score as explained above.
4. **Questionnaire:** Finally, the participant would be asked three open-ended questions to collect qualitative data on how they found the Study, to help contextualise the results:
    a. What part(s) of the information provided for each example did you find most helpful in understanding how the model makes predictions, and why?
    b. What part(s) of the information provided for each example did you find unhelpful and/or confusing in understanding how the model makes predictions, and why?
    c. What additional information would you like to have seen to improve your understanding of the model's predictions?

3.3.2 Survey Development

The Study was conducted using an online survey instrument to collect participant responses, which was developed using the platform Qualtrics. This was the preferred approach as it allowed the study to be conducted online and shared via web URL to participants, making it easier and quicker to recruit participants and collect responses. Further details on the survey are included in Appendix C.

The online survey was structured to support the format of the study procedure as set out in 3.3.1. Section 1 covered the pre-requisite information for subjects to decide whether to participate, including:

- **Participant Information Sheet**: Overview of what the Study involved, what was expected of the participant, and what would be done with the data collected

- **Participant Consent Form**: Setting out the terms the user would be agreeing to by participating in the Study, and capturing their explicit consent to participate.

At the start of the study, the survey was set up with logic to randomly assign each participant into either the treatment or control group, while also ensuring even allocation of participants across each group.

Section 2 of the survey was designed to support the Training Phase of the study, as described in 3.3.1. 12 pet listing examples were selected at random, 3 per predicted class, with image, text and tabular LIME outputs generated for each example using the customised implementation code explained in 3.2.1. Mock-ups were then designed of how the pet listing details and LIME outputs would be displayed to participants. These displays were produced for each example using the data visualisation software Microsoft Power BI, allowing for significant automation of the production process and quick refreshing when new examples were chosen.

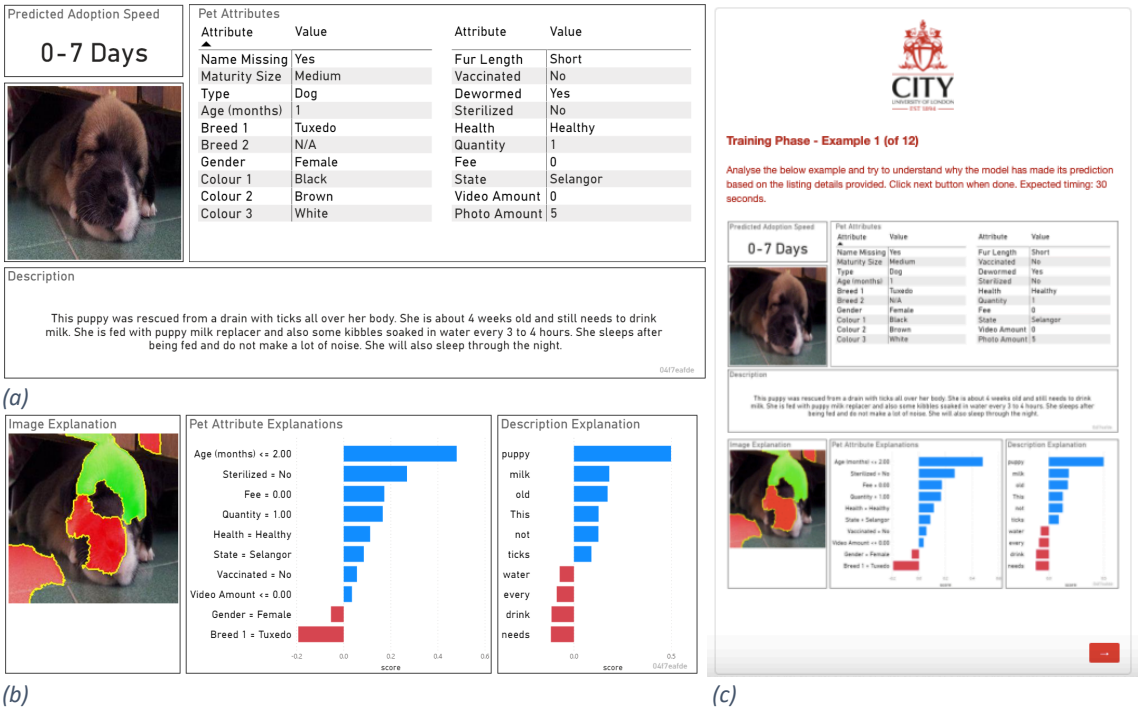


*Figure 10: Examples of output displays used in the **Training Phase** of the study. (a) pet listing display produced in Power BI; (b) LIME outputs display produced in Power BI; (c) displays embedded as images into the online study to be presented to participants. Treatment participants were presented with displays (a) and (b) in the study; Control participants were only shown display (a).*

These were then embedded as images into the survey to be presented to participants – providing a compromise between using informative graphical explanations while keeping

these static for simplicity, as explained in 2.3. Conditional logic was added into the survey flow to present participants with the relevant displays depending on whether they were in the treatment or control groups. Figure 10 shows examples of these displays and how they were embedded into the survey.

Section 3 of the survey aligned to the Prediction Phase of the study. 10 new examples were selected at random with LIME outputs produced and image displays embedded into the survey, as per the Training Phase. However, the displays omitted the model's prediction as these had to be estimated by the participant. Instead, prediction options were placed below the embedded image for participants to select their estimated adoption speed, in order to collect these data in the Study, as explained in 3.3.1. Figure 11 shows examples of the output display used, and how this was embedded into the survey.

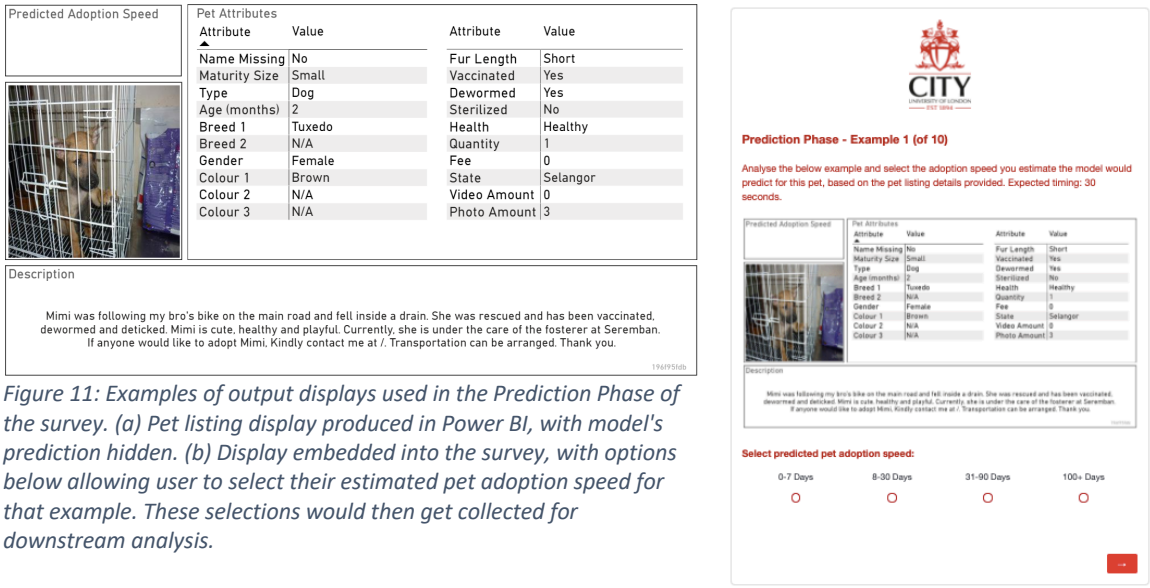


Figure 11: Examples of output displays used in the Prediction Phase of the survey. (a) Pet listing display produced in Power BI, with model's prediction hidden. (b) Display embedded into the survey, with options below allowing user to select their estimated pet adoption speed for that example. These selections would then get collected for downstream analysis.

The final section of the survey was aligned to the Questionnaire part of the study, designed to capture qualitative data on how the participant found the survey and what information they found useful. The section was set up with the Questionnaire questions outlined in 3.3.1, and free-text fields to capture participants responses, as illustrated in Figure 12.

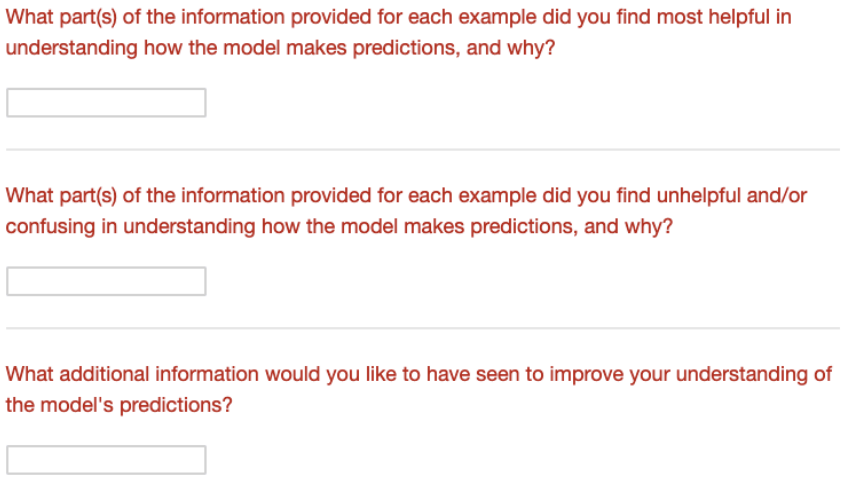


Figure 12: Layout out of questions in the survey relating to the Questionnaire part of the study

3.3.3 Conducting Study

Once the online survey was developed, the study could be set in motion. Subjects were recruited to participate using a two-pronged approach.

1. The survey was first shared with friends and family. These were eligible to participate since there were no pre-qualification criteria in place – the research question intends to test the effect of LIME explanations on users in general, rather than any niche group of specialists.
2. Once the first avenue had been exhausted, remaining subjects were then recruited through the paid platform Prolific, in order to arrive at the required number of 70 participants.

This two-wave approach was implemented in order to keep recruitment costs as low as possible, by minimising the number of subjects recruited via the paid platform Prolific. It took approximately two weeks to carry out this process and collect sufficient responses to perform the analysis.

In total, approximately 120 responses were collected, of which several were discarded due to poor quality, missing details, or because they had been completed too quickly. Of the remainder, the ones that took the longest to complete were prioritised to arrive at the final 70 responses used in the study analysis (35 from each group) – the rationale being that these would contain higher quality data, since users had taken longer to think about their responses.

3.4 Analysing Study Responses

The analysis of the data collected in the study involved first obtaining descriptive statistics about each group and plotting these visually to gain insights into the differences between them. Then this was followed by performing a non-parametric t-test – since the data did not appear to be normally distributed – to assess whether there was a statistically significant difference between the two groups, and thereby statistically infer whether the explanations had improved user understanding [6]. Below are the key choices used for the t-test:

- **Hypothesis:** There is a statistically significant difference between the treatment and control weighted kappa scores (QWK)
- **Null Hypothesis:** There is no difference between the treatment and control QWK scores.
- **Test Statistic:** Mann-Whitney U Statistic (non-parametric) to measure the difference in mean ranks between the treatment and control.
- **P-value:** Probability of seeing a test statistic U at least as extreme as the one observed, if the null hypothesis is true.
- **Level of Significance:** P value needs to be less than 0.05 in order for the null hypothesis to be rejected, and the hypothesis to be accepted

With these choices in place, the first step was to calculate the QWK score for each participant across all their predictions in the study. The QWK scores were then compared across the two conditions, by applying the Mann-Whitney U test to calculate the p-value and decide whether to reject the null hypothesis.

In terms of the qualitative data collected in the study, this was used as context to help interpret and explain the result of the hypothesis testing, and other key insights gleaned from the quantitative data. It was also used to inform future work building on this research, by highlighting key improvements to make to the overall design of the study and survey instrument.

4 Results

This section sets out the main outputs and results obtained from each of the main phases in this project: Modelling; LIME Explanations; and User Study. Below are the notable headlines within each area:

- **Modelling:** The model produced to perform the pet adoption speed prediction task achieved a QWK evaluation score ranking in the top 20 of the Kaggle Competition leader-board from which the task was sourced. Even so, the model contained a significant level of noise with predictions quite spread out around the true labels.
- **LIME Explanations:** A custom implementation of the LIME package was developed to handle the multimodal nature of the model. The image and text explanations produced showed significant levels of noise and randomness, with the tabular explanations providing the strongest signals and patterns to understand the model's predictions.
- **User Study:** Applying a t-test on the data collected from the study revealed that LIME explanations in fact worsened users' understanding of the model relative to the control group, with a statistically significant p-value of 0.05.

The remainder of this section delves into each of these areas, providing more information on the key results and illustrating these with figures and graphical outputs.

4.1 Modelling

The multimodal AI model used in this project – explained in more detail in 3.1 – achieved a Quadratic Weighted Kappa (QWK) score on the test dataset of 0.4365, placing it in the top 20 submissions of the Kaggle competition leader-board for this problem, out of a total of 1788 entries. It is worth noting that this is not a fully like-for-like comparison, as the exact problem addressed in this project was simplified by reducing the number of prediction classes to 4, as explained in 3.1.2. While this partly explains the model's strong ranking on the leader-board – as it was learning a simplified task – this can still be considered a good reassurance of the model's performance as the problem remains a similar one.

Looking more closely at the model's predictions on the test set – as shown by the confusion matrix in Figure 13 – a few more insights emerge on the model's behaviour and performance. First, it can be seen that the model's predictions are quite noisy and spread out per class, with

less than half of instances in each class being correctly predicted; as reflected in the model's overall Accuracy score – i.e. proportion of correct predictions – of 0.3775. However, on closer inspection, we find that misclassifications are more concentrated around the true classes, and become less frequent as the prediction gets farther away from the mark. This is also reflected by the model's Root Mean Squared Error (RMSE) score of 1.007, showing that the model's predictions are on average approximately one class out from the true class. This suggests the model has learned some signal from the data, and accordingly performs better than random guessing which would yield an Accuracy score of 0.25.

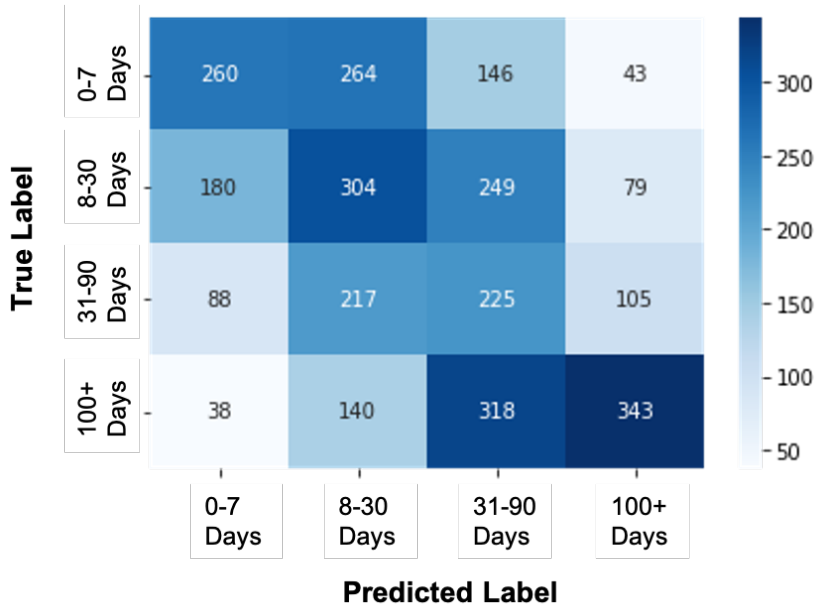


Figure 13: Confusion matrix showing the model's predictions on the test dataset, broken down by predicted versus true class labels. This indicates the spread of correct versus incorrect model predictions by class, offering more detailed insights on the model's performance.

4.2 LIME Explanations

One of the major results related to LIME in this project was the successful implementation of this technique on the multimodal pet adoption speed model – representing one of the main contributions of this work, as the open-source lime package only supports unimodal models out-of-the-box. Custom methods had to be written to allow this package to deal with the multimodal model, and the code for this is one of the main outputs from this stage of the analysis.

The main results found for each of the different LIME explanations – images, text, and tabular – are explained in more detail below.

4.2.1 Image Explanations

There did not seem to be much discernible insight emerging from LIME image explanations to help explain how images contributed to the model's predictions. Explanations tended to highlight the face and the body of pets in each image as contributing most positively or negatively to predictions – indicated by green or red highlighting respectively. This indicated the explanation provided some meaningful signal, in that it gave more weighting to the presence of the pet in the image over other background features. However, beyond this, it was not clear what specific features of a pet's appearance had a positive or negative effect or why, with similar pet features highlighted variably across different images – as illustrated in Figure 14.

Figure 14: Examples of LIME Image explanations where the model's predicted adoption speed is 0-7 Days. The explanations focus on the pet's features in the images. However, the same features appear to have variable impact (e.g. face), with no clear reason for this variability.

4.2.2 LIME Text Explanation

Text explanations provided more signal than image explanations, though the consistency of the explanations was still quite variable across examples. Text explanations are depicted as bar charts showing the words in the pet's description contributing most to the model's

prediction, as shown in Figure 16. Blue or red bars indicate whether the contribution was positive or negative respectively.

Analysing text explanations, it becomes apparent that certain words tend to positively contribute towards the model predicting shorter adoption speed. Notable examples include: *adoption, abandoned, rescued, kitten, puppy, milk, healthy, playing.* Interestingly as the model predicts longer adoption speeds, some of these same words come to take on a negative contribution to the prediction, such as *adoption, kitten,* and *puppy*. These words can therefore be seen to be the most predictive of adoption speed according to what the model has learned. Figure 15 and Figure 16 illustrates this pattern observed with the word *adoption* as an example.

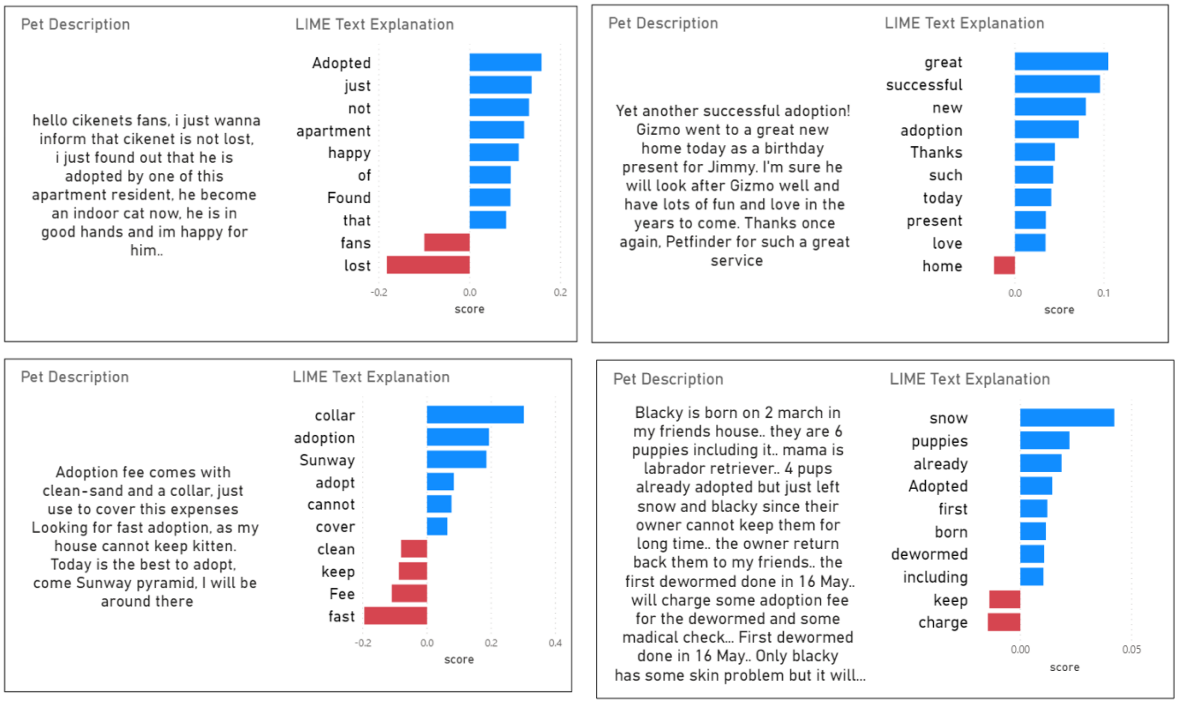


*Figure 15: Pet examples where predicted class is 0-7 days adoption speed. The recurrent appearance of the word **adoption** with a positive contribution, indicates that it is positively associated with this predicted class – i.e. its presence makes it more likely for the model to predict 0-7 days.*

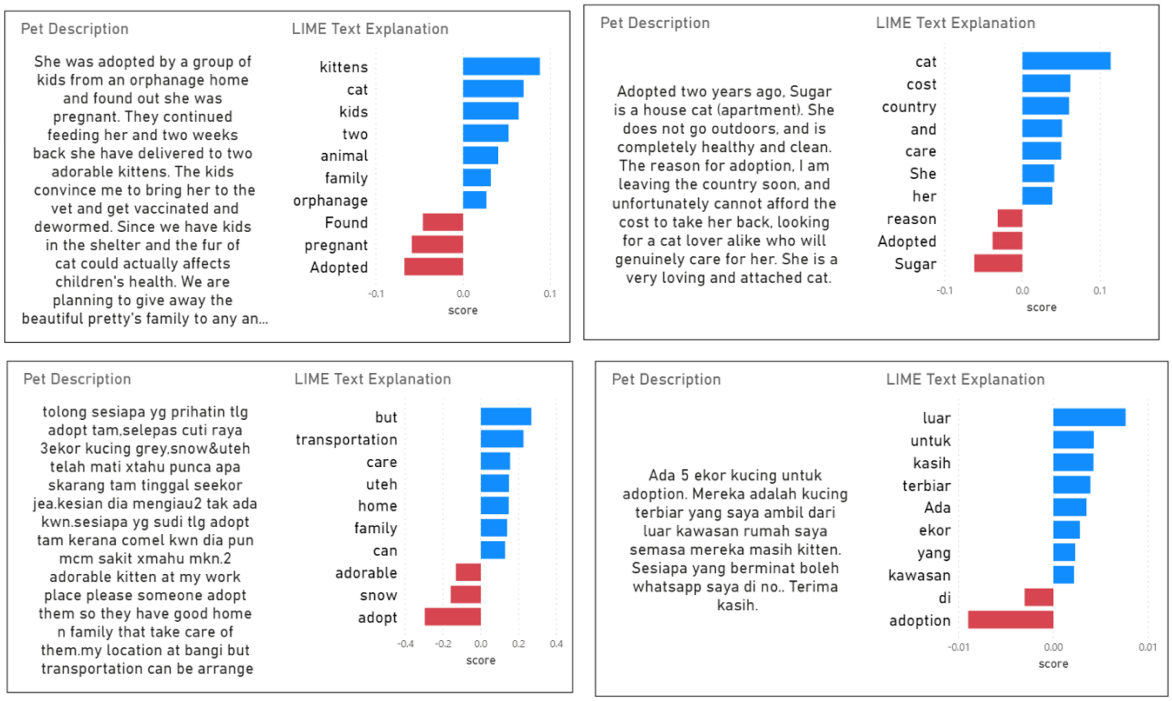


*Figure 16: Pet examples where predicted class is 100+ days, The recurrent appearance of the word **adoption** with a negative contribution, indicates that it is negatively associated with this predicted class – i.e. its presence makes it less likely for the model to predict 100+ days.*

4.2.3 LIME Tabular Explanations

Tabular explanations showed stronger patterns on how the various attributes of a pet (i.e. tabular data) affected the model's prediction behaviour. These data covered a range of characteristics about each pet including their *Type* (i.e. dog or cat), *Age*, *Gender, Breed, Colour,* and so on.

As per Text explanations, Tabular explanations are also shown as bar charts displaying the top attributes contributing most to the model's prediction, where the colour of the bar (blue or red) indicates whether the contribution is positive or negative; i.e. making the prediction more or less likely. Examples of these explanations are shown in Figure 17.

What emerged from analysing these explanations was a group of attributes which correlated with the predicted adoption speed based on what the model had learned. These included: *Age, Sterilisation, Quantity* (i.e. number of pets up for adoption in the listing), *Adoption Fee, Gender,* and *Breed* among others.

As shown in Figure 17 and Figure 18, when the model predicted 0-7 days adoption speed, this was positively associated with: *Age < 2 Months; Sterilised = No; Quantity = 1; Fee = 0;* and *Gender = Male*, and negatively associated with: *Gender = Female, Fee > 0, Age > 3*

Months. Conversely, when the model predicted 100+ days adoption speed, this was positively associated with: *Age > 3 Months, Sterilised = Not sure, Gender = Female,* and negatively associated with: *Quantity = 1, Fee = 0, Sterilised = No, and Gender = Male.*

Translating this into plain English, this suggests the model bases low adoption speeds on pets either being very young, not sterilised, free (no adoption fee), male, or the sole pet included in the listing. If pets are older, possibly sterilised, female, part of a multi-pet listing, or come with an adoption fee, then the model is more likely to predict a longer adoption speed.

Interestingly, when the pet in question had a rarer breed – e.g. Golden Retriever or Cocker Spaniel for a dog, or Oriental Tabby for a cat – this tended to contribute very positively to 0-7 Day adoption speeds and negatively to 100+ Day adoption. On the contrary, the presence of a more common breed – e.g. Tuxedo for a dog, or Common Long Hair for a cat – tended to have the opposite effect. This suggests, in addition to young age, rare breed is also a highly sought after attribute when adopting pets, based on what the model has learned.

4.3 User Study

One of the main products from the user study included the online survey instrument for users to participate in the study. More details on this survey can be found in Appendix C.

This survey was developed on the Qualtrics platform using the City University licence, and was designed with the relevant branch logic to handle the end-to-end study procedure for both Treatment and Control subjects, as explained in more detail in 3.3.2.

The other main product was the dataset collected for the study using the online survey. This was collected over a 2-week period – relatively quickly thanks to the ease of online dissemination of the survey to participants – and amounted to 125 participant responses overall. Of these, several were discarded due to poor quality, and 70 responses – 35 from each group – were taken forwards for analysis, as per the requirements of the study design, as set out in 3.3.1. The dataset included both quantitative and qualitative data for analysis, and the key results of each analysis component are set out below.

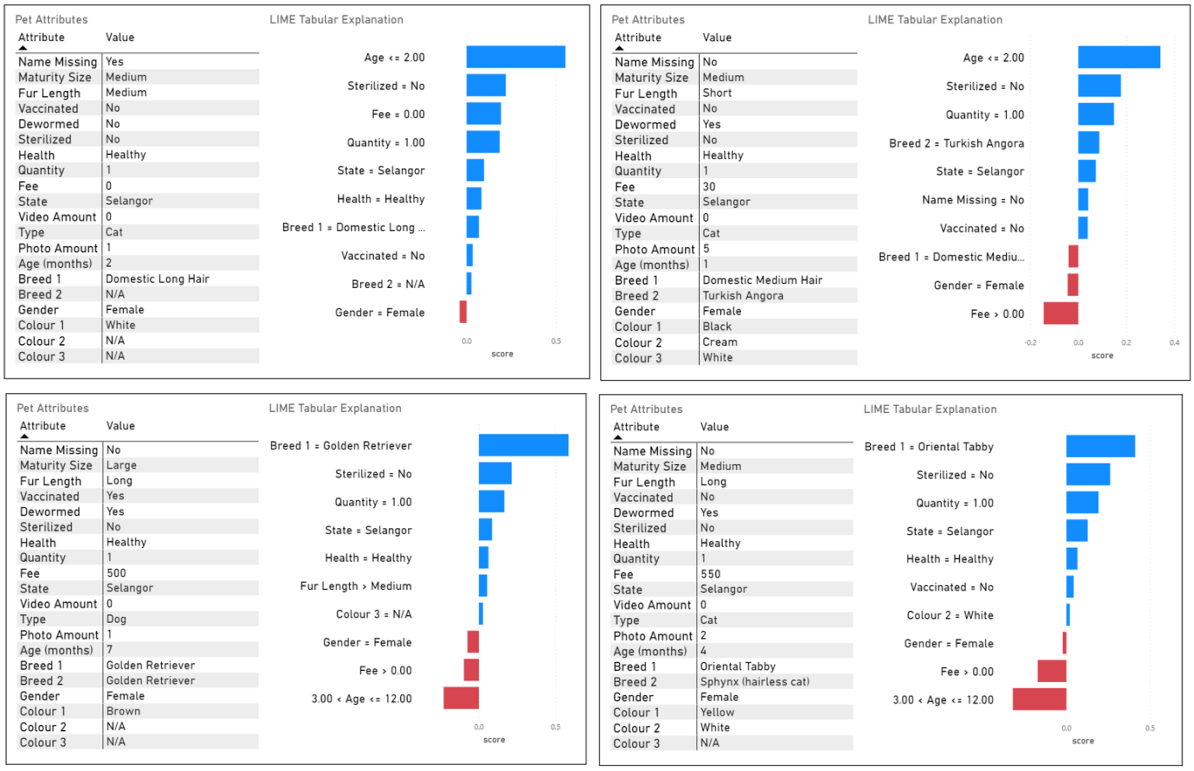


*Figure 17: Pet attribute examples, where the model predicted **0-7 Days** adoption speed, along with corresponding LIME tabular explanations. These examples indicate that pets that are **less than 2 months old, not sterilised, male, and free (no charge)** contribute positively towards this prediction category.*

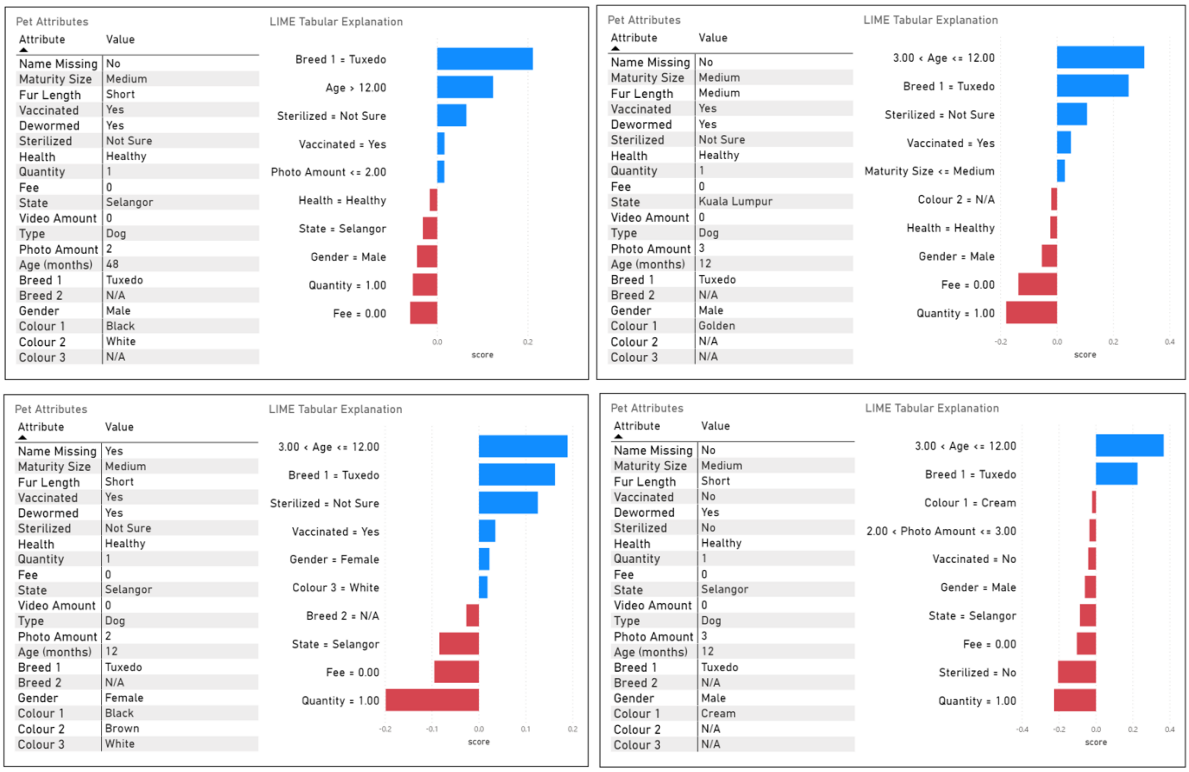


*Figure 18: Pet attribute examples, where the model predicted **100+ Days** adoption speed, along with corresponding LIME tabular explanations. These examples indicate that pets that are **more than 3 months old, (maybe) sterilised, and female** contribute positively to this prediction category.*

4.3.1 Quantitative Results

The quantitative data consisted of subjects' overall weighted kappa scores (QWK) across their predictions in the study (i.e. QWK scores per subject). As explained in more detail in 3.4, a t-test was run to assess whether the Treatment QWK scores were greater than those of the Control, with sufficient statistical significance – thereby inferring whether the explanations enhanced user understanding.

First, descriptive statistics were obtained on the control and treatment groups to gain insights into the key differences between these. These results are summarised in Table 2 and Figure 19 below.

Condition	Count	Mean	STD	Min	25%	50%	75%	Max
Treatment	35	0.302	0.299	-0.311	0.057	0.238	0.563	0.820
Control	35	0.417	0.212	-0.154	0.253	0.480	0.590	0.711

Table 2: Descriptive statistics for the treatment and control groups, showing: the sample size of each condition (count), their mean and standard deviation, their min and max values, and 25th, 50th and 75th percentiles

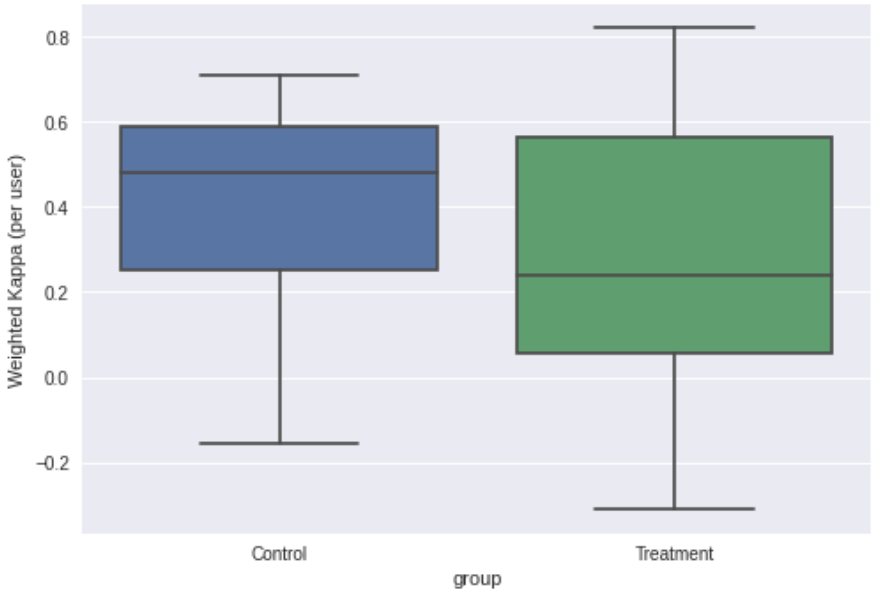


Figure 19: Box plot visually representing the descriptive statistics in Table 2, for the control and treatment conditions.

From the above, it can be seen that – counter-intuitively – the Control values appear to be higher and less spread out than those of the Treatment condition, with a higher interquartile range and mean value, and narrower standard deviation. This suggests control subjects tended to have a better understanding of the model than those shown the LIME explanations.

Next, statistical inference was performed on the data, by conducting a t-test to confirm whether this observed difference in the populations was statistically significant. As shown by

the histograms in Figure 20, the data did not appear to be normally distributed – with the treatment group showing a bimodal distribution, and the control group a left skewed distribution – and so a non-parametric t-test was applied; namely the Mann-Whitney U test [6].

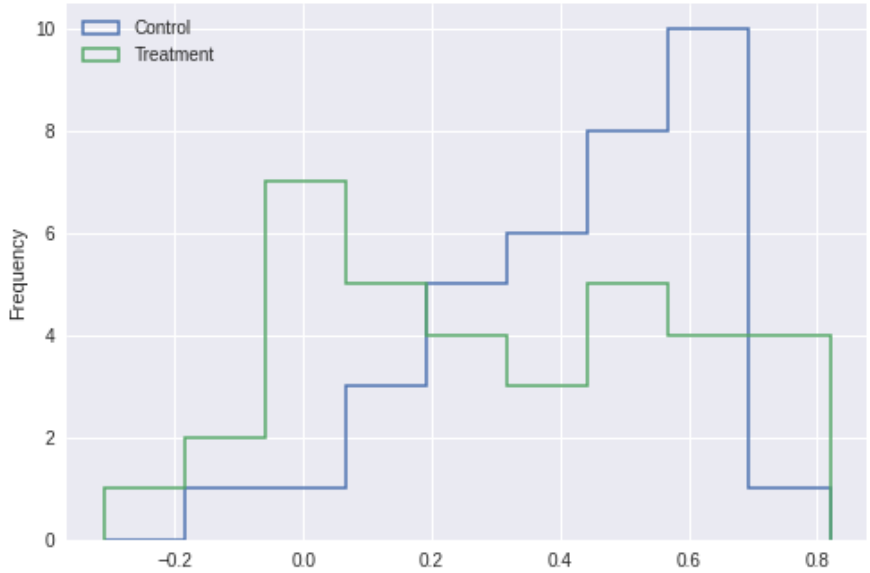


Figure 20: Histograms showing the distribution of the Control and Treatment populations

From conducting the t-test, the relevant p-value came out as 0.05, achieving the minimum level of significance to reject the null hypothesis that the two populations are the same; i.e. the positive difference in the control compared to the treatment is statistically significant. In other words, we can statistically infer from the study that exposing users to the explanations devised in this project worsens their understanding of the model's predictions.

Figure 21 sheds more light on this, by showing the prediction accuracy of control versus treatment subjects per each of the 10 prediction examples in the study. Notably, the Control group's accuracy can be seen to be markedly higher on several predictions, specifically predictions 5, 2 and 1.

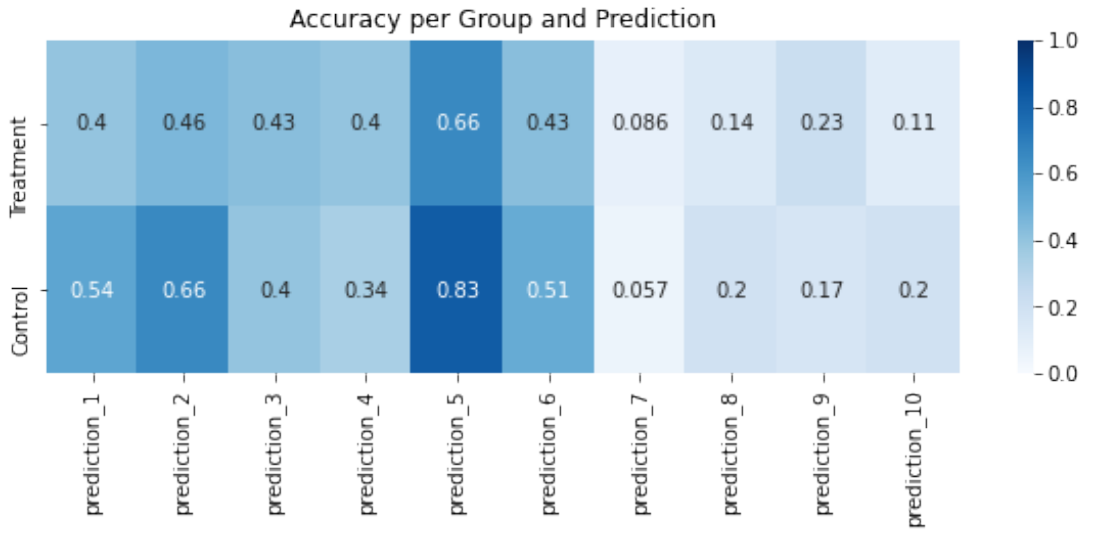


Figure 21: Accuracy of participants' predictions in the study by group, and per each of the 10 prediction questions in the study

4.3.2 Qualitative Results

As explained in more detail in 3.3.1, the qualitative data collected consisted of free-text responses to the following open-ended questions:

1. What part(s) of the information provided for each example did you find most helpful in understanding how the model makes predictions, and why?
2. What part(s) of the information provided for each example did you find unhelpful and/or confusing in understanding how the model makes predictions, and why?
3. What additional information would you like to have seen to improve your understanding of the model's predictions?

The aim here was to capture more open-ended feedback from participants to help contextualise and more easily interpret the quantitative results observed. These qualitative data revealed a number of key insights. Firstly, it was notable that 61% of participants cited the *Age* attribute as the most helpful piece of information in helping their understanding. Interestingly, this varied significantly across subject groups, with 69% of control subjects citing *Age* compared to 54% of treatment subjects – suggesting that it was more helpful for control subjects, perhaps because it was the signal that stood out to them most in the absence of LIME explanations.

Another curious observation was the variable mention across the subject groups of *Pet Attributes* as the most helpful piece of information. 25% of control subjects cited this, compared to only 2% in the treatment group, suggesting that the LIME explanations – which

only treatment subjects were exposed to – perhaps captured more of their attention, than the *Pet Attributes* information which was shown to all participants.

Focusing on participants' views on what they found most unhelpful, no clear pattern emerged amongst the control group. However, within the treatment group, certain recurrent views were noted. Specifically, a significant proportion of treatment subjects found both LIME Image and Text explanations to be unhelpful – 28% and 20% respectively – with several subjects citing the seeming noise of these explanations. This aligns with the observed randomness of these explanations as noted in 4.2.

There was also mention of certain design elements of the LIME explanations being unhelpful and requiring more information, with one participant commenting: 'Image explanation requires some form of scale. Attribute and description scores needs supporting information to give weight to the scores.'

In terms of additional information subjects would like to have seen, some asked for additional images of the pet to be shown as part of each example. Another request that came up multiple times was asking for more details on how the algorithm itself worked, to help with users' understanding of the model's behaviour. Finally, a few subjects mentioned the ability to view and compare information across multiple pet examples, to more easily see how predictions varied in line with explanations.

5 Discussion & Evaluation

This chapter discusses the results obtained in the project in more detail, analysing them in the context of the project objectives to determine whether these were adequately met, and identify the main findings to draw from the results. It then goes on to provide an evaluation of the project as a whole, summarising the main conclusions and contributions while also offering reflections on key lessons learned and elements of the project which are worth revisiting. Through these reflections, suggestions will then be put forward on directions for future work based on what has been achieved in this project.

5.1 Key Objectives Review

This section discusses and interprets in greater detail the results outlined in chapter 4, within the context of the project's main research question and key objectives, as well as the broader

research landscape where relevant. Below is a recap of the research question and project objectives used to answer the question:

Research Question:

> **How can we develop explanations for multi-modal AI models to enhance user understanding?**

Project Objectives:

1. **Modelling:** Producing a multimodal AI model achieving an acceptable performance threshold
2. **Explanations:** Applying LIME Explanations to each data type processed by the model, in a way that is readily understandable to users.
3. **User Evaluation:** Conducting a user study to empirically assess the effect of LIME explanations on users' understanding of the model's behaviour, thereby answering the research question.

In the remainder of this section, each of the objectives are reviewed in light of the key results observed in the project, with a view to highlighting to what extent they were met and could have been done better.

5.1.1 Modelling

As outlined in 1.3, the test criteria for this objective was that the model should rank within the top 100 submissions on the leader-board of the Kaggle competition from which the problem had been sourced. Based on this, the objective seems to have been met since the model ranked in the top 20 submissions of the leader-board.

However, even so, the model's performance was still quite low in absolute terms. As seen in 4.1, the model's misclassifications were quite spread out – albeit around the true adoption speeds – with an overall accuracy of 0.3775. Therefore, the model still showed significant noisiness in its predictions, in spite of its ranking.

This was reflected in the observed results for the LIME explanations – especially image and text explanations – where the same features were shown to have inconsistent impact on the model's predictions over different pet examples; something which was also corroborated by participants' comments in the user study. What this indicates is that perhaps a different

multimodal problem and dataset should have been chosen altogether. The fact that the model ranked highly against a leader-board consisting of over 1500 entries, and yet still contained significant noise, points to the inherent intractability and noise in the dataset itself, which does not lend itself well to the prediction task at hand.

The conclusion to draw here, therefore, is that the model and dataset used may not have been the best choices for this project after all, given the noisy predictions which impacted downstream analysis, especially the usefulness of the LIME explanations to users. To avoid this, the test criteria for the modelling objective could have been amended by stipulating that the model should achieve a minimum level of absolute performance, regardless of its relative performance against other published results. This would have perhaps resulted in better modelling results.

In addition, it is also worth noting that better results may have also been achieved through trying other modelling approaches. Particularly an end-to-end neural network model learning more informative joint multimodal representations from the data, and using these to perform more accurate predictions.

5.1.2 LIME Explanations

The test criteria here was that the explanations should be easy to understand for users. This seems to have been met in terms of the design of the LIME outputs for each data type, as study participants appear to have found them intelligible and easy to understand, based on their qualitative feedback in the study.

However, where the test criteria does not appear to have been met is in the actual usefulness of the LIME explanations to help study participants better understand the model's predictions. In this sense, while the explanations were easy to interpret, they were not easy to extract meaningful insight from. Indeed, as explained in 4.3.2, a significant proportion of treatment subjects in the study found the image and text explanations to be 'confusing' and 'random'. This may explain why the study revealed that the LIME explanations in fact worsen users' understanding of the model's predictions.

As explained in the above section 5.1.1, this randomness in the explanations is probably due to the noisy predictions of the model itself which in turn is likely the result of the dataset and problem being inherently noisy. However, this may have also in part been due to the lack of

experimentation with the hyperparameters of the LIME algorithm. As seen in 2.2.2, LIME is susceptible to instability in its explanations, and trial and error is required to discover the optimal configuration of key parameters such as the kernel size for data sampling, which heavily influence the explanations produced. This was not done in the project, with the hyperparameters left in their default settings, and so would need to be incorporated into the analysis pipeline in future.

5.1.3 User Evaluation

Based on the test criteria for this objective – to determine whether LIME explanations have a statistically significant effect on users' understanding, and thereby answer the research question – this objective was clearly met, and this phase of the work was an unequivocal success.

What the results revealed from analysing the quantitative data in the user study, is that LIME explanations did have a statistically significant effect, but not in the way that was expected or intended. Instead of positively improving users' understanding of the model's predictions, the analysis revealed that LIME explanations worsened user understanding, with a p-value for this hypothesis of 0.05 indicating statistical significance. This result clearly answered the research question in the negative, concluding that the LIME explanations used in this project do not enhance – but in fact harm – users' understanding.

The only aspect of the study that may require revisiting is the chosen sample size of 70 subjects. As explained in 3.3.1, this was based on the expected study parameters to achieve the required level of power of 0.8 for the study; specifically a mean Quadratic weighted kappa score of 0.7 +/- 0.3 for the treatment group, and 0.5 for the control group. In reality, these values were actually 0.3 +/- 0.3 and 0.4 respectively based on the data collected from the study. This would require a sample size of 282 to achieve the level of power of 0.8 – raising a question mark on the validity of the statistical results inferred from the study. This would need to be addressed, if the study were to be repeated.

5.2 Key Conclusions

The main conclusion to note is that the LIME explanations used had a negative effect on users' understanding of how the multimodal model studied in this project makes predictions. These explanations were intentionally designed to enhance understanding, which raises

questions on why they had the opposite effect, as borne out in the data collected through the user study. While there was mention in the study of design improvements required, the clear message that emerged was the randomness in the LIME explanations which 'confused' participants and led them to overlook key signals and identify patterns where there were none. This was evidenced by the significantly lower proportion of treatment subjects correctly identifying Age as the most predictive piece of information, and showing a stronger tendency to cite a greater number of factors as being helpful for prediction which seemed spurious and irrelevant.

The randomness in the explanations was likely the result of noise in the AI model used, in spite of comparing favourably against top submissions in the Kaggle competition from which the problem was sourced. While different modelling techniques could have been applied potentially yielding better results – for instance, using an end-to-end deep learning model for feature extraction and prediction – what this suggests is the intractability and noise inherent in the problem itself and the dataset used which did not lend itself well for the prediction task at hand. The implication to draw here is that the AI model in question needs to capture a certain level of signal in its predictions in order for any post-hoc explanations derived from it to be informative to human users and improve their understanding of its behaviour. Otherwise users simply see randomness in the explanations which can be confusing and in fact worsen their understanding.

Another potential cause is lack of stability in the LIME explanations themselves due perhaps to sub-optimal hyperparameters used when applying LIME to this problem. No trial-and-error manual tuning of key parameters such as kernel size was applied which could have resulted in more consistent and sensible explanation outputs. In addition, a certain level of instability may have been unavoidable since it is inherent in the LIME algorithm itself due to its reliance on locally permuted datasets for its explanations – datasets which have a certain stochasticity in how they are sampled, resulting in inconsistent explanations generated for the same local example, and also globally across multiple examples. The implication here is to consider post-hoc explanation techniques yielding more stable and globally consistent explanations, such as SHAP.

Even so, key contributions were achieved in this project, chief of which was the implementation of LIME on a multimodal model which required amending the source code of the lime open-source package to support such models, since they currently only support

unimodal models. No other previous work seems to have attempted this, addressing a clear gap in the literature which can form the basis for further research. Another notable contribution was the User Evaluation component of the project to assess the quality of the LIME explanations produced. The design and implementation of the user study to carry out the evaluation was effective and gave a clear and quantifiable answer. It could therefore be treated as a good template for assessing AI explanation techniques in general.

5.3 Project Evaluation & Reflections

5.3.1 Modelling

Reflecting on the Modelling phase, there are key changes that come to mind which could have had the potential to significantly improve the overall results of the project. The main lesson that emerged was the importance of the choice of AI problem and dataset. It is crucial to choose a problem with less noise in the data, where the state-of-the-art published solution achieves strong performance and exhibits clear signals in its prediction behaviour. Otherwise, as was observed in this work, any post-hoc explanation techniques used to try to shed light on the model will instead show randomness in its explanations which risks confusing rather than informing human users.

With hindsight, this lesson strongly points to the need to have chosen a different AI problem for this project, focusing instead on a well-established benchmark dataset within the Multimodal AI community. For instance, one alternative would be the Multimodal Opinion-level Sentiment Intensity (MOSI) [28] – consisting of annotated videos with accompanying audio and text for opinion and sentiment prediction tasks – which has been widely cited and has practical real-world applicability as per the original problem.

Focusing on the actual modelling techniques employed, the clear change here would have been to attempt a mid-level fusion approach to generate joint representations for training the model capturing the varying informativeness of each modality, rather than giving them equal weighting – as was the case in the early-fusion approach adopted in the project. Moreover, a full end-to-end deep learning model taking care of mid-level representation learning and prediction would have been the most promising avenue to explore further here.

5.3.2 AI Explanations

As established in this work, the LIME explanations used proved confusing to users and resulted in a negative impact on users' understanding of the model. It is therefore clear that this aspect of the project requires revisiting, to address how the explanations could have been rendered more insightful and less 'random'. As noted previously, this is also tied with the modelling since the level of the noise in the model also feeds into the explanations; if these are faithful as they should be. However, leaving to one side the modelling, a few key recommendations come to mind to address any instability present in the explanations.

First it may have been worth adopting a more stable post-hoc explanation technique from the outset such as SHAP. Unlike LIME, SHAP provides theoretical guarantees that feature contributions on the prediction are assigned fairly, and can also be aggregated to give a global view of the model's key drivers overall. This stability and global consistency could have helped users to more easily detect any signals present in the model's predictions, to inform their understanding. Secondly, rigorous manual tuning could have been applied on LIME to make the explanations more stable and reliable – something which was omitted from this aspect of the analysis.

5.3.3 User Evaluation

As noted previously, the User Evaluation component of this work was considered the most successful part of the project, with the designed study procedure and online survey instrument allowing the necessary data to be captured quickly to provide an unambiguous answer to the main research question.

That said, some changes and suggestions are worth mentioning, specifically the question around the sample size for the study. It would appear that the sample size may have needed to be significantly larger than the one actually used, on account of the study parameters deviating from the estimates on which the original sample size estimate was based. This raises questions around the validity of the results obtained and would need to be revisited in future work.

One final suggestion relates to the design itself of the instrument used to perform the study. While the online survey was deemed a success, the content presented to participants was displayed in a static format. It did not provide any interactivity allowing users to delve into

specific data items or patterns of interest, see more data examples than the ones hardcoded into the survey, or make comparisons on the fly across different examples to aid their understanding – all things that were requested by participants in the survey feedback. It would make sense therefore to develop an interactive online application to conduct the study in future, providing all the requested additional functionality.

5.4 Future Work

Finally it is worth outlining some avenues for further research based on the work achieved in this project. The first suggestion would be to repeat the project again, implementing the lessons learned and changes recommended in 5.3, to see whether more favourable results would be obtained. In particular, opting for a different multimodal AI problem, employing an end-to-end NN modelling approach, and switching to a more stable post-hoc explainer such as SHAP, may well yield enhanced user understanding of the multimodal model in question, as was the initial hope for this work. It would also be interesting to run an ablation study, conducting several variants of the analysis where only one of these changes is implemented at a time in order to compare results and identify what elements were specifically responsible for the negative effect on user understanding seen in this work.

Another suggestion relates to adapting the internal mechanics of the LIME algorithm to make it more fit-for-purpose for multimodal models. Specifically, in the way that it generates permuted samples to train the local surrogate models and generate explanations. New sampling methods could be explored for multimodal models which are more holistic, involving taking samples covering all modalities simultaneously, rather than sampling from one modality at a time and holding the others fixed – the approach used in this work to repurpose the unimodal focus of LIME's current implementation. Challenges will have to be overcome around how to come up with sensible samples given the heterogeneity of the different modalities, but this may well result in more informative explanations which take account of the complementarity and interdependencies of the different modalities in how they influence the model's predictions.

References

[1] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C., Parikh, D. and Batra, D., 2016. VQA: Visual Question Answering. *International Journal of Computer Vision*, 123(1), pp.4-31.

[2] Baltrusaitis, T., Ahuja, C. and Morency, L., 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp.423-443.

[3] Beheti, P., 2021. *Introduction To Multimodal Deep Learning*. [online] Medium. Available at: <https://heartbeat.fritz.ai/introduction-to-multimodal-deep-learning-630b259f9291> [Accessed 3 January 2021].

[4] Bowman, S., Angeli, G., Potts, C. and Manning, C., 2015. A large annotated corpus for learning natural language inference. In: *Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pp.632–642.

[5] Cao, Y., Long, M., Wang, J., Yang, Q. and Yu, P., 2016. Deep Visual-Semantic Hashing for Cross-Modal Retrieval. In: *KDD*.

[6] Dancey, C., 2020. *Statistics Without Maths For Psychology.* Slovakia: Pearson Education Limited.

[7] D'mello, S. and Kory, J., 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys*, 47(3), pp.1-36.

[8] Dodge, J., Liao, Q., Zhang, Y., Bellamy, R. and Dugan, C., 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. pp.275–285.

[9] Doshi-Velez, F. and Kim, B., 2021. *Towards A Rigorous Science Of Interpretable Machine Learning*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1702.08608v2> [Accessed 3 January 2021].

[10] Frome, A., Corrado, G., Schlens, J., Bengio, S., Dean, J., Ranzato, M. and Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In: *Advances In Neural Information Processing Systems, NIPS.*

[11] Kaggle.com. 2021. *Petfinder.My Adoption Prediction | Kaggle*. [online] Available at: <https://www.kaggle.com/c/petfinder-adoption-prediction> [Accessed 3 January 2021].

[12] Kaggle.com. 2021. *Petfinder.My Adoption Prediction Data | Kaggle*. [online] Available at: <https://www.kaggle.com/c/petfinder-adoption-prediction/data> [Accessed 3 January 2021].

[13] Kanehira, A., Takemoto, K., Inayoshi, S. and Harada, T., 2019. Multimodal explanations by predicting counterfactuality in video. In: *CVPR.*

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*. pp.3149–3157.

[15] Kulesza, T., Burnett, M. and Stumpf, S., 2015. Principles of explanatory debugging to personalize interactive machine learning. In: *Intelligent User Interfaces (IUI*).

[16] Lipton, Z., 2018. The Mythos of Model Interpretability. *Queue*, 16(3), pp.31-57.

[17] Liu, S., Kailkhura, B., Loveland, D. and Han, Y., 2019. Generative Counterfactual Introspection for Explainable Deep Learning. In: *IEEE Global Conference on Signal and Information Processing*.

[18] Lundberg, S. and Lee, S., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*.

[19] Lundberg, S., Erion, G. and Lee, S., 2018. *Consistent Individualized Feature Attribution For Tree Ensembles*. [online] Arxiv.org. Available at: <https://arxiv.org/pdf/1802.03888.pdf> [Accessed 3 January 2021].

[20] Martens, D. and Provost, F., 2014. Explaining Data-Driven Document Classifications. *MIS Quarterly*, 38(1), pp.73-99.

[21] Molnar, C., 2019. *Interpretable Machine Learning. A Guide For Making Black Box Models Explainable*. [online] Available at: <https://christophm.github.io/interpretable-ml-book/> [Accessed 3 January 2021].

[22] Park, D., Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T. and Rohrbach, M., 2018. *Multimodal Explanations: Justifying Decisions And Pointing To The Evidence.* [online] arXiv.org. Available at: <https://arxiv.org/abs/1802.08129> [Accessed 3 January 2021].

[23] Ribeiro, M., Singh, S. and Guestrin, C., 2016. Why should I trust you?: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp.1135–1144.

[24] Rykes, 2021. *The Hitchhiker's Guide To The Petfinder Competition*. [online] Kaggle.com. Available at: <https://www.kaggle.com/c/petfinder-adoption-prediction/discussion/81597> [Accessed 3 January 2021].

[25] Vanbelle, S., 2014. A New Intepretation of the Weighted Kappa Coefficients. *Psychometrika*, 81(2), pp.399-410

[26] Wachter, S., Mittelstadt, B. and Russell, C., 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*.

[27] Williams, A., Nangia, N. and Bowman, S., 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp.1112–1122.

[28] Zadeh, A., Baltrusaitis, T. and Morency, L., 2016. *MOSI: Multimodal Corpus Of Sentiment Intensity And Subjectivity Analysis In Online Opinion Videos*. [online] Arxiv.org. Available at: <https://arxiv.org/pdf/1606.06259> [Accessed 3 January 2021].

Appendix A – Modelling Outputs

Training data

https://www.kaggle.com/c/petfinder-adoption-prediction/data

Script to pre-process training images

https://colab.research.google.com/drive/1YQ0LtNHzV2pxScidEbVxZQnMnAzgmCBQ?usp=sharing

Script to pre-process remainder of training data and train model

https://colab.research.google.com/drive/1YpKGyDlUZtOq0sl9ckd6rZw79o3yjT1b?usp=sharing

Model training outputs

NB: These were the modelling outputs used to produce the LIME explanations incorporated into the User Study

https://drive.google.com/drive/folders/1pLkqD47D9VkHd-tq-9KW_CNlbm-IAeCi?usp=sharing

Appendix B – LIME Outputs

Custom LIME Package

NB: This package contains custom LIME explainer methods inheriting from the explainer methods in the original lime package. These explainers have been amended to support the multimodal model used in this project

https://drive.google.com/drive/folders/1xdRIeQlgd7zOD57J_9DbKRrO7kmr2Gq7?usp=sharing

Script to generate LIME Explanations

https://colab.research.google.com/drive/1_-12Uw9zhLwTFOP_2bj192hp-4xRiGWi?usp=sharing

LIME Explanation Outputs

NB: LIME Explanation outputs used in the research project

https://drive.google.com/drive/folders/1-3kTCDeXoovF5lF3t3ZtzaMNotPhRUr8?usp=sharing

Appendix C – User Study Outputs

Survey Instrument

NB: Online survey instrument to participate in the User Study.

https://cityunilondon.eu.qualtrics.com/jfe/form/SV_07h8NgBpgVobj0x

Study Data Collected

Quantitative data: https://drive.google.com/file/d/1PcAY4LCDY6sem-fSwqCiebCHQUYBU0jd/view?usp=sharing

Qualitative data: https://drive.google.com/file/d/18Zn1l56YUHJeOxoYFeYvDirFKoE98Ask/view?usp=sharing

Script to conduct Statistical Analysis

NB: Script to conduct statistical hypothesis testing on the quantitative data collected from the study to answer the research question

https://colab.research.google.com/drive/1n7thJhs9XcIEBVaTVk2R4wzp0yJdQKQx?usp=sharing