國立政治大學資訊管理學系

碩士學位論文

機器學習可解釋技術在商業智慧中對使用者信任之影響

The Effect of Explanation on User Trust in Business Intelligence

指導教授: 林怡伶 博士

研究生: 侯亮宇 撰

中 華 民 國 110 年 7 月

DOI:10.6814/NCCU202101328

ACKNOWLEDGEMENT

To my family,

To participants, participate in the experiment to support my data collection,

To my advisor, Dr. Yi-Ling Lin giving me guidance to complete my research,

To the community that is exploring human-computer interaction.

DOI:10.6814/NCCU202101328

摘要

近年來機器學習引發了人工智慧 (Artificial Intelligence, AI) 應用的新趨勢。AI 被應用於越來越複雜的任務和領域中。然而，大多數 AI 模型都在黑盒(Black box)中運行，導致人們難以理解或是分辨機器的運作以及決策過程。目前，可解釋性人工智慧(Explainable Artificial Intelligence, XAI)，大多著重於底層演算法的解釋，並且集中於解釋圖形識別的結果。針對終端使用者的 XAI 應用則較多專注於支援醫療保健領域的人類決策，少有研究調查商業領域的 AI 應用程序如何與解釋性技術相結合。本研究以商業應用上終端使用者為中心為實際業務領域中運用 AI 技術提出了一個通用的解釋框架。該框架基於商業智慧(Business Intelligence, BI) 所開發，為終端使用者提供在機器學習不同階段的完整解釋。為了實踐我們的框架，我們在一個航空公司行李重量預測案例上應用了這個解釋性架構。最後，為衡量該框架實踐後的有效性，我們在 Amazon Mechanical Turk 上進行了實驗。我們的結果表明，使用解釋性框架的參與者對模型預測更有信心，並且更信任系統，更願意採用系統提供的建議。我們的研究使企業能夠擴展他們的商業智能，並結合這個解釋框架的不同階段，以提高機器學習技術在商業應用中的透明度和可靠性。

關鍵詞：*人機互動、機器學習、資訊視覺化、可解釋性人工智慧、信任*

DOI:10.6814/NCCU202101328

Abstract

Recently, machine learning has sparked a new trend in artificial intelligence (AI) applications. AI is applied to increasingly complex tasks and in many areas. Most AI models are running in a black box resulting in difficulty for understanding. From image recognition to sentiment analysis, XAI is used to support human decision-making in the healthcare domain, yet little research has been done to investigate how AI applications in the commercial domain can be integrated with explanatory techniques. This study proposes a generalized interpretative framework for end-user-centric applications in the business domain. The framework enables the provision of complete explanations to end users at different stages based on business intelligence. To validate our framework, we applied this explanatory framework in practice using an airline baggage weight prediction case. Finally, in order to measure the effectiveness of the framework in practice, we conducted an online experiment at Mturk. Our results show that participants who use the explanatory framework have more confidence in the model predictions, trust the system, and are more willing to adopt the recommendations provided by the system. Our research allows companies to extend their business intelligence and combine different stages of this explanatory framework to improve the transparency and reliability of machine learning technology in business applications.

Keyword*: Human computer interaction (HCI), machine learning, information visualization, trust, explainable artificial intelligence, XAI*

DOI:10.6814/NCCU202101328

Table of content

DOI:10.6814/NCCU202101328

DOI:10.6814/NCCU202101328

List of Tables

DOI:10.6814/NCCU202101328

List of Figures

DOI:10.6814/NCCU202101328

Chapter 1 Introduction

1-1 Background and Motivation

In the past few decades, various developments have been made in the field of artificial intelligence (AI). AI is now gradually entering the mainstream and provides tremendous support for human decision-making. The effectiveness of AI is limited by the inability of machines to explain its process and results to human users in various situations (Doshi-Velez & Kim, 2017). In the medical field, physicians are not clear about how an AI model predicts diabetes (Krause et al., 2016). In the business field, humans do not understand how the stock AI agents work; they often can only use historical back-testing to judge the pros and cons of the stock investment made by the agents (LeBaron, 2001). These AI applications may only be selected based on performance indicators, such as accuracy or precision, and not necessarily based on the interpretability. This is caused by the fact that we are using black-box models in AI.

When facing a complicated black box model, some users might "over-reliance on automation" and abuse the results (Parasuraman & Riley, 1997), while some might underestimate the power of model and concern about obsolescence, so called "ignorance or underutilization of automation" (Parasuraman & Riley, 1997). Both abandonment and misuse can cause serious problems. In order to use and understand AI correctly, users must trust the system appropriately (Barredo Arrieta et al., 2020a).

The prevailing belief is that explaining to users how the system works increases their trust in the system and reduces upset after seeing errors (Glass et al., 2008). Transparency and trust can be provided through verbal and visual means (Wang et al., 2016). Currently, many existing studies have focused on explaining the underlying algorithms. The designer may assume that users understand the decision-making process, or users had sufficient knowledge of machine learning models. However, the end user may not have sufficient background to understand how the algorithm works.

DOI:10.6814/NCCU202101328

The difference between the designer's hypothesis and the user's background may result in interpretive method not being able to effectively enhance the end user's trust in machine learning (Wang et al., 2019).

Many interpretive techniques focus on the interpretation of image recognition, such as the saliency map (Bach et al., 2015). These interpretation methods claim to have a good interpretation effect in many different fields (Ribeiro et al., 2016), but most of them only focus on the application in the medical field except for image recognition (Wang et al., 2019). These techniques are mostly designed for the machine learning developers. Despite of some techniques are end-user oriented, they are mostly used in the medical field for professionals (Vassiliades et al., 2021). We believe that there are many areas in the business domain where machine learning can bring value, but the lack of interpretation of predicted results may make it more difficult to apply the results in practice. Therefore, we apply explanatory techniques in the business domain and propose an explanatory framework based on BI. We explore whether this framework can enhance end-users' trust in the results of machine learning in the business domain and finally facilitate their decision making in the context of real business cases.

1-2 Research question

Some people refuse to use models to help them make decisions (Dawes, 1979; Dietvorst et al., 2015). There are several possible reasons for this situation: first, people may feel that they do not understand the model, including the information they rely on and how this information is used. For example, in a study on the use of machine learning in government departments, researchers noted that it was challenging to get organizations to accept the use of machine learning-based systems when the internal structure of the system could not be explained (Veale et al., 2018). Second, people may feel that the model does not use information in the right way or relies on incorrect input

DOI:10.6814/NCCU202101328

information (Veale et al., 2018). Third, people may be concerned that the models are operating in an unjust manner (Rudin, 2019). These reasons may lead to increased concerns about the use of models.

In previous approaches, interpretability techniques are mostly used in the fields of graphical recognition and medical interpretation. Business intelligence can provide different perspectives on complex data, including structured and unstructured data. The variety of data types can be used in different ways by different users such as sales, managers and employees. The combination of visual interaction systems and business model predictions can bring AI technology closer to the end-user, thereby improving overall operational efficiency and performance.

Therefore, we investigate whether the application of XAI in the business world can enhance user trust in the model. In order to measure user trust we refer to the trust calibration study proposed by Wang (N. Wang et al., 2016) where explanation and transparency lead to increased trust and team performance. On the other hand, the Davis et al.(Davis et al., 2020) study suggests that the focus on trust is somewhat narrow, leading the research community to stray from tried and true empirical methods, so we added confidence indicators and user acceptance. We considered this element because confidence and trust are important factors in the willingness of users to accept and use the process output (Pieters, 2011).

This study attempts to answer the following research questions:

- RQ1. Does the use of the XAI framework in the business field allow end-users to follow the predictions of machine learning results?
- RQ2. Does the application of XAI framework in the business field increase the confidence of users in adopting machine learning results?
- RQ3. Does the application of XAI framework in the business field increase end-users' trust in machine learning results?

DOI:10.6814/NCCU202101328

In this study, we begin with a review of the existing literature on different XAI technologies and articles that combine human-computer interaction and trust. Next, we propose a business intelligence-based AI interpretability framework that provides explanations to end-users in the business domain. The framework organizes the way in which explanations are provided to end-users in the past, and we have divided the explanations into three parts that designers can follow to provide explanations to user.

To evaluate the practical application of the explanatory framework proposed in this study in the business domain, we apply this framework to a real-world business data machine learning use case of predicting the weight of airline baggage to enable airlines to more accurately calculate fuel consumption and reduce the cost per flight. Finally, we deploy and conduct an online survey experiment at Amazon Mechanical Turk[1] to evaluate the effectiveness of the framework in real-world applications.

[1] Amazon Mechanical Turk is a crowdsourcing Internet marketplace where individuals or businesses can recruit participants to perform tasks that computers are not able to do.

DOI:10.6814/NCCU202101328

Chapter 2 Literature Review

In this chapter, we review the existing literature from three aspects. In the section 2-1, we present how business intelligence and decision support enable human understanding of analytics. In the section 2-2, we review the interpretability approach to AI and its advantages and limitations, and finally in the section 2-3, we focus on the interpretability of machine learning from the perspective of trust.

2-1 Business Intelligence

2-1-1 The Definition of Business Intelligence

Business intelligence (BI) refers to the use of data exploration, cloud computing, data analysis, and other technologies to interpret past sales data, operating costs, depreciation and amortization, and sales revenue data, as well as to convert the data into information, which can be provided to management as a reference "wisdom" for decision-making and judgment (Sautto, 2014). The development of modern business intelligence began in the mid-1980s, and is a further extension of Decision Support Systems (DSS), with the purpose of assisting corporate management in making decisions (Power, 2002). According to David Loshin's definition of BI, BI processes, technologies, and tools need to convert data into information, information into knowledge, and then knowledge into action plans that can gain company benefits (Saxena & Srinivasan, 2013). Technologies of BI include data warehousing, enterprise analysis tools, and content/knowledge management.

BI is arranged in the data warehouse for users to obtain real-time and dynamic high-value information anytime and anywhere from the data warehouse through a variety of online query, analysis and processing tools, data mining, and decision support systems to improve corporate decision-making. It is a mechanism to improve quality, improve performance, and achieve the ultimate goal of an enterprise (Chaudhuri et al., 2011).

DOI:10.6814/NCCU202101328

2-1-2 The Application of Business Intelligence

BI is supported by a corporate portal on the front end and a complete information integration system on the back end of the enterprise, allowing everyone from decision makers to grassroots employees to access important reference data for improvement; and to save a lot of form processes and enhance internal operational processes; creating high productivity and competitiveness. Business intelligence is based on building a data warehouse, integrating different operating system databases, and after cleaning, extracting, converting, and loading data from different sources and types, in a uniform form, Organized arranged in the data warehouse, users can obtain real-time and dynamic high-value information anytime and anywhere from the data warehouse through a variety of online query, analysis and processing tools, data mining, and decision support systems to improve the enterprise a mechanism for the quality of decision-making, improving performance, and achieving the ultimate goal of the enterprise (Dresner, 2001). BI has functions such as data management, data analysis, knowledge discovery, and enterprise optimization.

BI is an effective analysis mechanism that integrates the three elements of "management", "decision-making" and "information technology" (Gorchels, 2000). Companies must look at BI from a strategic perspective to understand its importance. In terms of application, BI has widely applied in the fields of customer relationship management, supply chain management, enterprise resource planning, or knowledge management that is well-known in the corporate world. It is the practical application of business intelligence (Shafiei & Sundaram, 2004).

2-1-3 The Tool in Business Intelligence

BI uses dashboards to compose data into charts and present the charts. While various charts are adopted in BI, this section focuses on the visualization tools for large

DOI:10.6814/NCCU202101328

volumes of data.

Treemaps (Bruls et al., 2000)**.** This method can be applied to large amounts of data to represent the data layers of each hierarchical structure in an iterative manner. Regardless of device resolution, analysts can always move to the next module to continue processing more detailed data at a lower level. Therefore, it can meet the large data volume standard. Since this method is based on the shape and volume estimates calculated based on one or more data factors, each time the data is changed, the image is completely redrawn for each currently visible hierarchical structure. Higher-level changes do not need to redraw the image because the analyst cannot see the data contained in it.

Circle Packing (Collins & Stephenson, 2003)**.** This method is a variant of Treemaps that uses torus. The inclusion in each circle represents the composition in the hierarchy: each branch of the tree is represented as a circle, and its sub-branch is represented as an inner circle. The area of each circle can also be used to represent any other value, such as a number or file size. Color can also be used to assign categories or represent another variable with a different shade. The main advantage of this method is that by using the classic Treemaps, we may be able to place and perceive more objects. Compared with Treemaps, this method can have better spatial applications.

Sunburst (Cawthon & Moere, 2007)**.** This illustrates the same type of data as the Treemaps. The highest level in the hierarchy is in the inner ring; the sub-levels are in the outer ring like the Treemaps. And because of this feature, you can use animation to modify the method to display data dynamics. It's easily perceptible by most humans.

Circular Network Diagram (Cawthon & Moere, 2007)**.** Data objects are placed around a circle and linked by curves according to their relevance. Correlation is usually measured using different line widths or color saturations.

Parallel Coordinates (Inselberg & Dimsdale, 1990)**.** The visual analysis of

DOI:10.6814/NCCU202101328

histones will be extended to multiple data factors for different topics. All data factors to be analyzed are placed on one axis, and the relative value of the data object is placed on the other axis. Each data object consists of a series of linked wires that indicate its position in other objects. This method allows us to use only a thick line on the screen to represent a single data object, and this method can meet the large amount of data.

Streamgraph (Kosara, 2016)**.** Streamgraph is a stacked surface diagram. A graph of displacement around the central axis, resulting in lowered and organic shapes. It shows the evolution of several groups of numerical variables. The area is usually displayed with the central axis as the center and the edges are rounded to create a flow shape.

2-1-4 The Challenge of Business Intelligence

BI is the "analytics role for decision makers", which allows decision makers to have a general understanding of a large amount of data obtained by the company through reports and charts, and then assists in decision making (Chen et al., 2012). Gartner has compiled a set of four categories of data analytics: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics (Burton et al., 2006). The first problem with BI is that it is still in the descriptive analytics, not in the diagnostic analytics (Deng & Chi, 2012). However, knowing what is happening is not very helpful for business decision making, it is more important to understand why it is happening. The second difficulty of BI: BI may not be real-time in its predictions and analysis of large amounts of data. The business world is highly competitive and there is no time for slow decisions. As today's business world becomes more competitive, companies need decision aids that can respond to market demands in real time. The next generation of analytics software with built-in algorithms has officially replaced BI, but users without a background in data analysis may not be able to turn algorithms into

DOI:10.6814/NCCU202101328

analytical tools. To solve problems that BI has not been able to solve in the past, we have incorporated the concept of XAI into the BI process. In the next section, we introduce explainable AI and its potential to develop and bring value to the BI architecture.

2-2 Explainable Artificial Intelligence (XAI)

2-2-1 The Reasons of XAI

The concept of XAI can be traced back to its first introduction by Swartout (1983) and Clancey (1983) in their expert system. It is gaining wide popularity in the field of ML interpretability in recent years (Carvalho et al., 2019). While AI is gradually entering the mainstream and provides tremendous support for human decision-making, the effectiveness of AI is limited by the inability of machines to explain its process and results to human users in various situations (Doshi-Velez & Kim, 2017).

In solving prediction problems, data scientists tend to focus on accuracy metrics (e.g., RMSE, Recall, etc.). High-accuracy models are often too complex, especially deep neural networks, which are often considered as black boxes because of the multiple hidden layers and the large number of non-linear weights, making it almost difficult to understand the relationship between the input and the resultant output and the decision making of the algorithm. Although accuracy is important, these metrics only provide partial information about the model, and the correlation and hidden information in the data cannot be expressed with accuracy. If we want to extend AI technology to more domains, we must try to understand how models make decisions and promote public trust in them. The following summarizes a few reasons why models need to be explained:

Verification of the system. To confirm the operation of the model, we make assumptions about the training capability of the model when constructing the model.

DOI:10.6814/NCCU202101328

For example, in a model that distinguishes between huskies and wolves, we expect that the model captures the facial features, contours, and body shapes of the animals in the picture. With the selection of training data, the model learned that as long as the background is white (snow), it is likely to be recognized as a wolf, otherwise, the background is not white, it is likely to be a husky. Before the explanation, even if the result is predicted correctly, we cannot know the relationship between model input and output. Sensitivity analysis and layer-wise relevance propagation (LRP) are commonly used explanation method which used heatmap to observe the correlation between each pixel and predicted result (Bach et al., 2015). These approaches are not limited to the model architecture, and can be presented in a visual way that is easy for humans to understand.

Improvement of the system. In order to improve the model, it is important to know the weaknesses of the model. The black box of the model makes the analysis more difficult. Through the explanation mechanism, it is possible to peek into the relationship between the predictions and the results of the model, making the relationship between them clear and making the potential risks of the model more visible, thus greatly increasing the possibility of avoiding errors and loopholes and improving the mastery of the model. In addition to avoiding potential risks, it is important to continuously improve its performance. Having good explanatory mechanisms, i.e., knowing the correlation between specific inputs and outputs, can provide direction in improving the models, even if they have the same performance, the features of interest vary between models (Lapuschkin et al., 2016). Through comparison, it is possible to identify the features that each model and architecture is good at or concerned with and to extract their strengths and exploit them.

Learning from the system. The AI system is trained with tens of millions of data, and it may observe potential data patterns that humans cannot find. When using an

DOI:10.6814/NCCU202101328

interpretable AI system, we can try to get knowledge from the AI system to gain new insights.

Compliance to legislation. Many areas of our lives are gradually being affected by artificial intelligence systems. For example, in terms of law, the distribution of responsibilities when the system makes wrong decisions has recently received increasing attention. Since relying on black box models may not find appropriate answers to these legal questions, future artificial intelligence systems have to be easier to interpret. Individual rights may become the driving force behind the increasing emphasis on artificial intelligence interpretation in regulations. For example, the "right to explanation" promoted in the new EU regulation GDPR allows users to request an explanation of her or his own algorithmic decisions (Goodman & Flaxman, 2017).

2-2-2 The Application of XAI

In the previous section, we explored the benefits of XAI for various purposes, including validating systems, improving systems, making them easier to learn, and responding to regulations. In order to put the theory into practice, in this section, we present practical applications of XAI. We detail how interpretative methods have been proposed and applied in machine learning, as well as introduce interpretative algorithms proposed in past research, and illustrate the use of global and local interpretations.

Making a Model Transparent. The transparent model itself has some degree of interpretability. A model is considered to be transparent if the model itself is understandable (McGovern et al., 2019). It has the properties of algorithmic transparency, decomposability, and model ability. For example, Linear Regression solves prediction problems and binary or multivariate classification problems (Hoffrage & Gigerenzer, 1998). Decision trees is a very common and excellent supervised learning algorithm, where the rules of internal nodes are trained to become hierarchical

DOI:10.6814/NCCU202101328

decision structures, making them easy to understand and highly interpretable (Quinlan, 1987). Rule-based learning refers to generating rules to characterize each model of the data to be learned from. Rules can take the form of simple conditional if-then rules, so they give a more understandable model (Langley & Simon, 1995).

Increasing Post-hoc explainability. When Machine learning models cannot provide explanations transparently, individual methods must be devised to provide explanations for the models. The purpose of post-hoc explainability techniques is to convey understandable information about the developed model that generates its predictions for any input. According to Barredo et al., (Barredo Arrieta et al., 2020b) the post-hoc explainability can be categorized into Model Specific techniques designed for application to any type of ML model and Model Agonistic models designed for specific ML models.

- *Model-specific*: The interpretation of a particular model is limited by the type of model. In order to obtain a specific type of interpretation method, only a specific model can be selected for the task. For example, if you want to obtain a tree-shaped decision diagram of a model task, you can only obtain it through a tree-shaped model. Such a restricted approach makes researchers turn their attention to unspecified model interpretation methods that can be applied to any model (Adadi & Berrada, 2018).
- *Model-agnostic*: Model-agnostic is not limited to a specific model, that is, model prediction and interpretation are two different parts. This type of interpretation method mostly uses post-hoc interpretation to analyze the relationship between prediction and model.

  According to different interpretation mechanisms, it can be divided into four types: (1) Visual explanation: Visualizing the operation mode of the neural unit of the deep neural network model.

DOI:10.6814/NCCU202101328

(2) Knowledge extraction, which transforms the internal representation of the model into an understandable form, observes the rules of the model's task, in addition to clarifying the mode of operation of the model, it can also explore new model.

(3) Feature relevance explanation: Change the input of features and the internal structure and parameters of the model, and observe the changes to the output results.

(4) Example-based: Observe a single example and provide possible explanations for the operation of the model from the predicted results (Adadi & Berrada, 2018).

Table 2-1 XAI methods category

Methods category	**Explanation**	**Definition**	**Algorithm**
Global interpretability	Global feature importance	Describe the feature weights used by the model and show the visualization of the feature weights	(Henelius et al., 2014; Lou et al., 2013; Nguyen et al., 2015)
Local interpretability	Local feature importance and saliency method	Show how the features of the instance contribute to the prediction of the model, presented in graphics or text	(Lundberg & Lee, 2017; Ribeiro et al., 2016; Simonyan et al., 2014)
Inspect counterfactual	Feature influence or relevance method	Show how the prediction changes corresponding to changes of a feature and often in a visualization format	(Adadi & Berrada, 2018; Apley & Zhu, 2020; Krause et al., 2016)
Example based	Prototypical or representative examples	Provide an example that is similar to the example and the record is the same as the forecast.	(Bien & Tibshirani, 2011; Koh & Liang, 2017)

Enhancing Global interpretability. Global interpretability provides an explanation for the entire operation mode and logic of the model. It is usually used in

DOI:10.6814/NCCU202101328

larger systems, such as weather operations, disease occurrences and other related factors that are very intricate and complex. A model that masters global weather conditions, with high probability It is better to grasp the overall situation and learn the patterns than just looking at a regional weather model. In order to have a more holistic perspective, the global interpretation model used to propose must be completely built under many restrictions, so it is limited by the trade-off between performance and interpretability. In practice, the realization of global interpretation is also more difficult. Global means a larger system, more variables and parameters, making it difficult to obtain an explanation that can fully describe the logic of the model operation. Therefore, local interpretation is a better way in reach in practice (Adadi & Berrada, 2018).

Enhancing Local Interpretability. Compared with the global explanation that explains the entire system, the regional explanation is a single predicted result, explaining the relationship between the result and the model. Interpretation methods include calculating the saliency map and heat map of the relationship between pixels and results; comparing the influence of different features on the model prediction results; observing the gradient trend between different labels, etc.

- *Local interpretable model-agnostic explanations (LIME)*：LIME was proposed by Ribeiro et al. (2016) and it is Model agnostic. The assumption behind it is that we can basically understand some models as long as they are not too long and too complicated. For example, in image recognition, the linear regression model of super-pixel may be understandable (we accept the explanation that "a large number of pictures with certain characteristics is an image of something"); in text classification, bag- The linear regression model of of-words is understandable ("a lot of words containing certain keywords belong to a certain category"). LIME is trying to use this understandable model to locally fit the model and prediction results you want to explain. LIME can interpret its "black box" separately for any

DOI:10.6814/NCCU202101328

given supervised learning model. Local interpretation means that LIME give a true local interpretation around or near the interpreted observations. LIME has a rich open-source API and can be used in R and Python, so it has a huge user base, with nearly 8k stars and 2k forks in its GitHub repository.

- *Shapley additive explanations (SHAP):* The Shapley value was proposed by Lloyd Shapley, a professor at the University of California, Los Angeles, to solve the problem of contribution and income distribution in cooperative games. In N-person cooperation, the contribution of individual members is different, and the income distribution should also be different, and the ideal ·distribution method is that the contribution and the income are equal.

2-2-3 The Challenge of XAI

In the field of machine learning, the more complex the model is, the more difficult it is to interpret. The rule usually states that the more accurate the model, the more complex the model, and the more difficult it is to interpret its output. As shown in Figure 2-1. The simplest model is linear regression. In statistics, a linear regression is a type of regression analysis that models the relationship between one or more independent variables and a strain using a least squares function called a linear regression equation. The corresponding relationship makes the output result easy to interpret; the decision tree is composed of a decision diagram and possible results to create a plan to reach the goal. Decision tree is a special tree structure used to assist decision making. Each node of this structure can present decision-making results, and interpretation is relatively easy.

DOI:10.6814/NCCU202101328

Figure 2-1. Machine Learning interpretability, "accuracy/interpretability" trade-off rule.

In machine learning and deep learning, deep learning is an algorithm based on the representation of data for learning. Observations can be represented in a variety of ways, such as a vector of intensity values per pixel, or more abstractly as a series of edges, regions of a particular shape, and so on. And it is easier to learn tasks from real examples using some specific representations. The benefit of deep learning is to replace the manual acquisition of features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction. Such algorithms often contain many hidden layers that are not easily understood by the user in the decision-making process, and even with XAI methods, it is often possible to obtain only the weight of each feature on the output. These interpretations may provide explanations to model designers and further improve the models, but they may not be easily linked to the end-users.

Although the authors of the XAI approach claim that these techniques can be applied in multiple domains, it still has limitations. As mentioned earlier, XAI method are mostly used for graph recognition and medical applications, but rarely for business intelligence. Business intelligence refers to the use of modern data warehousing technologies, online analytical processing technologies, data exploration, and data

DOI:10.6814/NCCU202101328

presentation technologies for data analysis to realize business value. The concept of business intelligence was understood by the public after being popularized by Dresner Howard (2001). At that time, business intelligence was defined as a type of technology and its application consisting of data warehousing, query reports, data analysis, data exploration, data backup and recovery, etc., for the purpose of helping companies make decisions.

2-3 Trust

In this section, we discuss the trust of humans in machines. Trust plays an important role in the decision of end users to use automated systems. In the past, explanation mechanisms often mentioned how machine learning explained how to convince users of machine judgments. With the latest developments in the field of artificial intelligence, more and more decision-making tasks are entrusted to the system. This section cuts into the application of XAI in business intelligence from the perspective of trust.

2-3-1 Trust in Computer Sciences

A person's level of trust in someone or something can determine how well that person accepts things. Trust is the main reason for acceptance. Trust is important in many relationships, such as interactions between people. Trust can also define the way people interact with technology. Trust is an essential component of various human interactions (Groom & Nass, 2007), which enables people to act under uncertainty and risk of negative consequences. It enables people to act under uncertainty and risk of negative consequences. In real-world human-robot interactions with high risk and high uncertainty, distrust reduces people's willingness to accept robot-generated information and follow robot advice, thus limiting the potential benefits of robotic systems (Freedy et al., 2007).

In computer science, trust is a widely used term whose definition varies between

DOI:10.6814/NCCU202101328

researchers and applications. Trust is an important component of the Semantic Web vision, where new problems and applications of trust are being studied. This paper provides an overview of existing trust research in computer science and the Semantic Web.

2-3-2 Measuring Human Computer Trust

In general, trust is a complex concept that is somewhat related but not exactly similar to confidence. In the Madsen & Gregor (2000) study, human-computer trust is defined as the degree to which users have confidence in and are willing to act upon the recommendations, actions, and decisions of an artificial intelligence decision aid. This concept was modified from McAllister (1995) and was chosen as the clearest and most complete definition of the concept of Human Computer Trust. It includes users' confidence in the system and their willingness to act on the system's decisions and recommendations.

The relationship between explainability and trust has been discussed in several recent papers (Lipton, 2018; Ribeiro et al., 2016). Related to our research, and the inspiration for our experimental design, Cai et al. (2019) ran a series of experiments with randomized human subjects and found that any of the different explanatory mechanisms had a significant increase in human trust, and that certain graphical elements did improve people's ability to detect when the model was wrong.

For example, Yang et al. (Yang et al., 2020) investigated the effect of exemplar-based machine learning classifier interpretation on appropriate trust for end users. The effects of spatial layout and visual representation were explored in a participant face-to-face user study. Changes in user trust over time were also observed. Results show that each explanation increases user trust in the classifier, and that the combination of explanation, human, and classification algorithm yields better decisions than if the human and classification algorithm were separate.

DOI:10.6814/NCCU202101328

The relationship between system performance and user trust in automated systems (Yu et al., 2016), ubiquitous computing systems (Kay et al., 2015), and recommender systems (Panniello et al., 2016) has been examined in several studies from the HCI community.

Poursabzi Sandeh et al. (2021) used house price prediction as a background, and they conducted a joint experiment in which subjects were presented with randomly generated house features and asked each subject the total price of the object. These models differed in their stated accuracy, the size of the fictitious training dataset, the number of features, and several other attributes. The authors estimate the impact of each attribute by fitting hierarchical linear models and find that one is usually most concerned with the size of the training dataset, the origin of the algorithm, and the accuracy of the statements to trust the decision of the model less about the transparency of the model or the relevance of the training data.

Zhang et al. (Zhang et al., 2020) conducted a case study of AI-assisted decision making where humans and AI alone have comparable performance and explored whether features revealing case-specific model information can calibrate trust and improve the joint performance of humans and AI, specifically they investigated the effect of displaying confidence scores and local explanations for specific predictions. They also highlight the use of local explanations in AI-assisted decision-making scenarios to calibrate human trust in AI.

Based on a review of past literature, trust is an important factor in the effectiveness of human-computer interaction systems and frameworks, so our framework performs a comprehensive measure of its trust effectiveness.

DOI:10.6814/NCCU202101328

Chapter 3 Research methodology

Since there are few studies exploring the feasibility of XAI in the business field, in order to verify the effectiveness of the XAI method in the business field, we propose a design framework and present it in conjunction with business intelligence tools. Then, we use a set of business cases to apply machine learning techniques combined with an interpretation framework to provide the final result. Users use and discuss its effects. Whether the trust in machine learning can be improved through interpretation.

3-1 Theoretical Background

In the previous chapter, we discussed different interpretation methods and algorithms, which can provide different degrees of interpretation to the predicted results. This study starts by distinguishing the explanation needed by different users. Different groups of audiences seek different goals for interpretability in ML models. Two goals are intertwined: the need to understand the model and compliance. In the case of Data scientist and developers, the explanation is needed to ensure or improve the performance of the model, to do further research, or to propose new features (Barredo Arrieta et al., 2020b). Domain experts, such as medical doctors, need to trust the decision basis of the model itself to assist in further treatment. Finally, the users who are affected by the model decisions need to know where they stand in order to verify that future decisions are feasible. They are also the recipients of the explanatory framework proposed in this study. From an AI research perspective, the recent review by Nunes and Jannach summarizes several explanatory purposes (Nunes & Jannach, 2017). Through transparent explanations, users can see some aspect of the internal state or function of an AI system. When using AI as a decision aid, users seek to use explanations to improve their decisions. In order to focus on enhancing the trust of end users, we looked at how to design XAI for users, as mentioned by Wang et al. (D. Wang et al., 2019) to discuss human decision-making with the user as the center. Liao et al.

DOI:10.6814/NCCU202101328

(2020) through interviews with colleagues in different departments of IBM's views on XAI, sorted out the question users want to know about XAI. In a Microsoft paper proposed by Amershi et al. (2019) they offered 18 universal design guidelines for human-to-human interaction. And it has been verified in interactive application scenarios. They divided the design of interpretable AI into four sections, namely “Initially”, “During interaction”, “When wrong” and “Over time”. At the initial stage, users should clearly understand what the system can do, and know when the system might go wrong; during interaction, add time, contextually relevant and other elements; then you need to let users know when the system goes wrong through invocation, dismissal, correction, and Draw the appearance of the service has reduced the user's doubts; the last is to explain the timeout part, learn the user's behavior, and adopt the user's feedback. Putnam & Conati (2019) discussed when and whether it is necessary to explain its user modeling technology to students in the Intelligent Tutoring System (ITS). And believes that incorporating the interpretation into ITS is beneficial to the overall impact.

Outside the ML field, Miller (2019) and others explored the space of user needs, using a problem-driven framework to explain. They proposed that the explanation is "the answer to the question". At the same time, they put forward the type of explanatory question and the reasoning that needs to be answered.

Table 3-1 Type of explanatory question and the reasoning that needs to be answered by Tim Miller

Question	Reasoning	Description
What	Associative	Given the observed events, the reasons for which unobserved events may have occurred.
How	Interventionist	Simulate the changes in the situation to see if the event is still occurring.
Why	Counterfactual	Simulate other causes and see if the event still occurs.

DOI:10.6814/NCCU202101328

3-2 Framework Development

Inspired by the prior work, we use the framework proposed by Miller (2019) as the framework base, distinguished by “what”, “how” and “why” then combined with the results of Liao et al. (2020) interviews with end-users at IBM. The framework of this study is based on BI. It is a descriptive analysis that allows decision makers to have a general understanding of the data obtained through reports and charts to assist in decision making. (Chen et al., 2012). We define the original explanation of BI as the What stage, and do the descriptive analysis, while the further model prediction results are presented in the "How" and "Why" stages by combining the explanatory algorithms. Based on the architecture proposed by Miller (2019), we divide what users want to know about What, How, Why, three directions:

What: This stage represents descriptive analysis as in the original BI system, including input and output data, data type, appearance, selected data characteristics, data used by the system, etc.

How: This stage represents how the model performs calculations and decisions and provides explanations in conjunction with the global interpreter.

Why: This stage represents why this prediction result is calculated. The weights of the features affected in the Model calculation are presented.

According to the study of user trust by Davis et al. (2020) and Yang et al. (2020), presenting the performance of the machine department or the confidence index can help improve the user's trust in the model, so we decided to add performance to the framework and put it in “How” stage.

This explanatory framework covers multiple aspects of using machine learning models for prediction to help complex associations of data in the business domain. First, the pre-processing of data can be presented in the What stage, such as the type of data, composition, amount of data, and the selected features. In the model prediction stage,

DOI:10.6814/NCCU202101328

users use the accuracy to judge the performance of the model based on their past experience, and we combine this stage to present the performance metrics. Finally, we produce the results, and we provide the reasons for the results through post hoc explanatory algorithms. From these three steps, we help the user to build a complete understanding rather than a single item of explanation.

We then standardized the explanation content that users wanted to know based on the end-user interviews conducted by Liao et al. (2020). In the What stage, we divided the data input and output in order to make the user's clearer about the data type. In the How stage, to make users understand the system logic, we explain it through Global explanation. Also, according to the research of Zhang (2020), the accuracy can improve the user's trust, so the performance is included. Finally, to make users understand the output results, we provide explanations by post hoc. The discussion for each type of interpretability requirement is finally divided into five blocks to provide explanations to users, as shown in Table 3-3.

DOI:10.6814/NCCU202101328

Table 3-2 Six different oriented explanations, and pre-set the problems that users may have.

Category		What user want to know	How	Tool
What	Input/data	What kind of data does the system learn from? What are the sample-size? What data is the system NOT using? How much data is the system trained on?	Shows the data composition, fields, quantity, selected features, and how much data to use for training.	Histogram, Pie chart, Treemap, Circular packing
	Output	What Kind of output does the system give? What is the scope of the system's capability?	Presents Numerical data, possible ranges, and values given in the past using average values.	
How	Global explanation	System overall logic? How does it weight different feature? What kind of the algorithm is used?	Shows what kind of algorithm is used, the logic behind it (XGBoost), how he gives each feature weight, and uses LIME, SHAP and other explanatory techniques to present global feature importance	PCA
	Performance	How accurate are the predictions? How often does the system make mistake? In what situations is the system likely to be correct? What kind of mistakes if the system likely to make?	Through the model accuracy (confidence level), the distance between the decision and the actual value.	Confidence level, accuracy
Why	Local explanation	What feature of this instance leads to the system's prediction	Use SHAP to present the contribution of each feature in the prediction result (Local interpretability)	Tornado Plot, Saliency Map, Description

DOI:10.6814/NCCU202101328

Input/data. The training data helps users to properly evaluate the motivation of using the model. Users can evaluate the model by querying sample size, potential bias, and whether data are missing. The presentation and transparency of the raw data allows users to better understand the limitations of the model.

Output. Understanding output is usually an overlooked aspect of XAI's algorithmic work, but we think we need to provide such an explanation. We describe the functional scope of system prediction, how the output data will affect other systems, etc., and finally through the past prediction methods The numerical value is compared to explain to the user.

Performance. Performance indicators are considered to be an effective tool for enhancing user trust in many studies (Yang et al., 2020; Zhang et al., 2020), and some studies believe that performance indicators may deter users. Some people also think that small differences between these indicators will not change the way users interact (Liao et al., 2020). We suppose that performance indicators can help users understand the limitations of AI and make it feasible to answer "whether performance is sufficient to meet the requirements of...".

How (Global). Provide a global explanation of how AI makes decisions, which can not only help users properly evaluate system functions, but also build a mental model to better interact with the system or improve the system. This type of demand is particularly prominent in: "Which of these attributes the company cares about is the most important...". In order to answer the How question, XAI algorithms usually use ranking features or decision trees. We remind users which models are used in this prediction and how their feature weights are adjusted.

Why (Local). Many current XAI algorithms focus on the "why" problem. We have noticed that the challenge of algorithm interpretation is that the comparison results are often not explicitly used in the model. These observations once again demonstrate the

DOI:10.6814/NCCU202101328

benefits of interactive interpretation, allowing users to clearly refer to the comparison results and make follow-up comments. According to our constructed explanatory framework Table 3-3, we can see that SHAP can provide multiple presentations in the why stage and the application of the suite is more complete than other existing algorithms (Lundberg & Lee, 2017). Therefore, we choose to use SHAP for the local interpretation of the weights of each feature in the prediction results.

This research proposes a framework based on BI combined with machine learning interpretation and integrates the information that end users want to know, so that data and training output results can be systematically presented according to needs at different stages. Next, we describe the interpretive tools that can be used in this interpretive framework, and we consolidate studies that apply these tools. We also detail the strengths and weaknesses of these methods and the contexts in which they are suitable for use.

According to the classification of Vilone & Longo (2020) on the output format, visual explanation is probably the most natural way to communicate and is a very attractive way to explain. Visual explanation can also be used to illustrate the internal functions of a model through graphical tools. For example, heatmaps can be used to highlight specific areas of an image or specific words of text by using different colors, which mainly affect the reasoning process of the model. For humans, another intuitive form of interpretation is textual interpretation, i.e., natural language statements that can be expressed in writing or orally. Therefore, we divide the final presentation into graphical and textual interpretations. The following is a list of visualization techniques applied in different situations. Basic bar charts, pie charts can be used to present raw data, including discounted data with time series.

DOI:10.6814/NCCU202101328

Table 3-3 Visualization of framework

		Graph											Text		
		Histogram & Bar	*Pie Chart*	*TreeMap*	*Circle Packing*	*Sunburst*	*Tornado Plot*	*Saliency Map*	*Waterfall Plot*	*Image Plot*	*Scatter*	*Partial Dependence*	*Text Plot*	*Confidence*	*Text Description*
What	(van Wijk et al, 1999)		●	●	●										
	(Robertson et al., 2008)							●			●				
	(Borkin et al., 2013)		●	●				●			●				
	(Yur & Vasil, 2013)	●		●	●	●					●				
	(Pandey et al., 2014)	●									●				●
	(Allen, 2018)	●	●	●											●
	(Xia et al., 2020)	●													●
How	(Yagoda & Gillan, 2012)													●	●
	(Desai et al., 2013)													●	●
	(Kim et al., 2019)	●	●												●
	(Zhang et al., 2020)													●	●
	(Yang et al., 2020)									●				●	●
Why	(Ribeiro et al., 2016)									●		●	●	●	
	(Lundberg & Lee, 2017)						●	●	●	●	●	●	●		
	(Bach et al., 2015)							●					●		
	(Samek et al., 2017)							●		●			●		
	(J. Wang et al., 2018)							●		●	●	●			
	(Alber et al., 2019)							●		●					

Graphical interpretation

Histogram & Bar Chart (Pizer et al., 1987). The histogram mainly presents the results of data distribution, while the bar graph presents the size of each data group. The horizontal variables of the histogram are "numerical continuous variables", and the bar chart is "class-based discrete variables". As for the "spacing" of group distances, the histogram groups are connected together and there is no spacing between them; the bar chart has spaced between groups.

DOI:10.6814/NCCU202101328

- Advantages: Can be used at any time.
- Disadvantages: easy to lose information.

Pie chart. A pie chart is a circular statistical chart divided into several sectors, used to describe the relative relationship between volume, frequency, or percentage.

- Advantages: The pie chart can effectively display the information. Especially when you want to show the proportion of a large sector in the whole, rather than compare, this method is very effective. When the proportion of the pie chart reaches 25% or 50% of the parent body, the display purpose can be well achieved.
- Disadvantages: It is difficult to compare different sector sizes in a pie chart, or to compare data between different pie charts.

Tree Map (Bruls et al., 2000). This method can be applied to large amounts of data to represent the data layers of each hierarchical structure in an iterative manner.

- Advantage: Since this method is based on shape and volume estimates calculated from one or more data factors, it is easy to meet a large number of data presentation standards.
- Disadvantage: Every time the data changes, the image needs to be redrawn according to the structure.

Circle packing (Collins & Stephenson, 2003). This method is a variant of Treemaps that uses torus.

- Advantage: By using classic Treemaps, it is possible to place and perceive more objects.
- Disadvantage: Same as tree map, every time the data changes, the image needs to be redrawn according to the structure.

Sunburst (Cawthon & Moere, 2007). The sunburst chart can express the level and attribution relationship of the data on the basis of the pie chart showing the proportion relationship, and can clearly express the data with the parent-child hierarchical structure

DOI:10.6814/NCCU202101328

type.

- Advantages: It has the advantages of a pie chart, and it can clearly show the relationship between data levels.
- Disadvantages: Not suitable for data display with too many data classifications, no negative values, and no zero values.

Tornadoes Plot (Molnar, 2019). To present localized explanations, vertical bar graphs of tornadoes can be used for attribution lists (Ribeiro et al., 2016). A tornado diagram is a popular tool for describing the sensitivity of the result to changes in the selected variable. It shows the effect on the output of changing each input variable once and leaving all other input variables at their initial values.

- Advantages: The tornado graph graphically shows the correlation between the changes in the model input and the distribution of the results.
- Disadvantages: Overemphasized the extreme value of the change of sensitive elements, but may ignore the difference in the possibility of various extreme values.

Saliency Map (Itti et al., 1998). A data visualization technology that displays the absolute amount of a phenomenon in the form of color in a two-dimensional space. The change of color may be through hue or intensity, to provide readers with obvious visual cues.

- Advantage: Image prominence is an important visual feature of the image, which can reflect the importance of the human eye to each area of the image.
- Disadvantage: susceptible to interference by noise.

Waterfall Plot (Gillespie, 2012). A waterfall chart is a three-dimensional graph in which several curves of data, usually spectrums, are displayed simultaneously.

- Advantage: This kind of chart uses the combination of absolute value and relative value, and is suitable for expressing the quantitative relationship between several specific values.

DOI:10.6814/NCCU202101328

- Disadvantage: Since the vertical graph in the waterfall chart represents a single feature along the horizontal line, the waterfall chart is suitable for the presentation of fewer features.

Scatter (Touchette et al., 1985). The scatter diagram is used to analyze the relationship between a pair of parameters, and the paired data is plotted on the X-Y diagram to find the relationship between the two.

- Advantages: Intuitively reflect the data concentration situation, and assist in the fitting of discrete data linear regression and other curve predictive fits.
- Disadvantages: less applicable scenarios.

Partial Dependence (Friedman & Meulman, 2003). Partial dependence diagrams were proposed by Friedman (2001) to understand the relationship between a feature in the model and the mean of the predicted target y, assuming that each feature is independent and presented in a visualized manner. Partial dependence diagrams have been used to visualize the variation of feature attribution across feature values (Krause et al., 2016). The interpretation of images and graphics can use technologies such as saliency heatmaps or pixel analysis. These visualization technologies support comparative interpretation and counterfactual reasoning by comparing different attributes or understanding the relationship between factors.

- Advantage: Partial dependence diagrams are easy to generate computationally and intuitive to understand.
- Disadvantage: It can only present at most two features in relation to y at the same time, and more than three dimensions cannot be presented by current techniques. It also has a strong assumption of feature independence, which can lead to bias in the estimation process if there is correlation between features.

DOI:10.6814/NCCU202101328

Textual interpretation

Text Plot. The text plot is a combination of text and heatmap. Passing a single instance to the text map get the importance of each tag overlaid on the original text corresponding to the tag (Lundberg & Lee, 2017). In the context of SHAP, in the context of sentiment analysis models, red corresponds to more positive comments, and blue corresponds to more negative comments. The importance values returned for the text model are usually hierarchical and follow the structure of the text. Non-linear interactions between marker groups are usually saved and can be used in the drawing process.

- Advantage: Use Heatmap feature to render on text, easy to understand.
- Disadvantage: An area of input text needs to be segmented by a good interpreter, otherwise the text map is regarded as a unit.

Confidence. According to Zhang et al., (Zhang et al., 2020) research results, showing confidence scores improves trust calibration and increases people's willingness to rely on AI predictions in high confidence situations.

- Advantage: The confidence level and accuracy of the model can effectively enhance the trust of users.
- Disadvantage: Under the explanation that only provides the confidence level, the user is not aware that it may cause the user to over-trust

Text Description. Giving user feedback through text narration can enhance the user's trust in the machine and provide a reference basis for the user's judgment when the machine may have errors (Desai et al., 2013).

- Advantage: Textual interpretation is an intuitive form of interpretation, which can be expressed in written or verbal natural language.
- Disadvantage: Compared with other interpretability methods, text interpretation is less used. Structured plain text explanations may not be effective in enhancing

DOI:10.6814/NCCU202101328

user confidence.

3-3 Framework Evaluation

To verify the feasibility and effectiveness of this explanatory framework, a case study of an airline company that applying machine learning to predict baggage weight would be used in this research. The case study would apply the explanatory framework to provide business intelligence-based explanations to end-users. Finally, we conducted a between-subject experiment to verify the effectiveness of the framework and whether user trust is enhanced by the explanation. The next two chapters introduce the case study and the experimental design respectively.

DOI:10.6814/NCCU202101328

Chapter 4 Case Study

4-1 Business Question

Aircraft fuel consumption is an important issue in airline operations planning and analysis, according to the National Airspace System (NAS) (Wong et al., 2009). Although recent fuel prices represent only a small portion of aircraft operating costs (approximately 16-22%), they are still a significant expense for airlines and general aviation operators (Trani et al., 2004). The weight of cargo and baggage is one of the most important factors affecting fuel (Hong & Zhang, 2010), and the inability to effectively predict the weight of passengers' checked baggage in advance can lead to aircraft carrying excess fuel for this purpose. The current airline solution to this problem is to use historical data to obtain an average of checked baggage weights for the route to estimate the fuel for that flight.

The flight control department prepares the flight plan according to the flight regulations, company policies, flight routes, weather conditions, destinations, standby stations, and the minimum amount of fuel required for the flight, etc. The weight of passengers, cargo, and passenger baggage are considered in Figure 4-2. Checked baggage is calculated based on one 20kg baggage per person, multiplied by the number of passengers booked on that flight. The weight of the passenger's checked baggage is added to the number and weight of the required luggage containers. The weight of the cargo is provided by the cargo handling unit at the airport.

Based on this premise, this case tries to develop a suitable method for estimating passengers' checked baggage using an artificial neural network approach to improve the accuracy of business units in fuel calculation.

DOI:10.6814/NCCU202101328

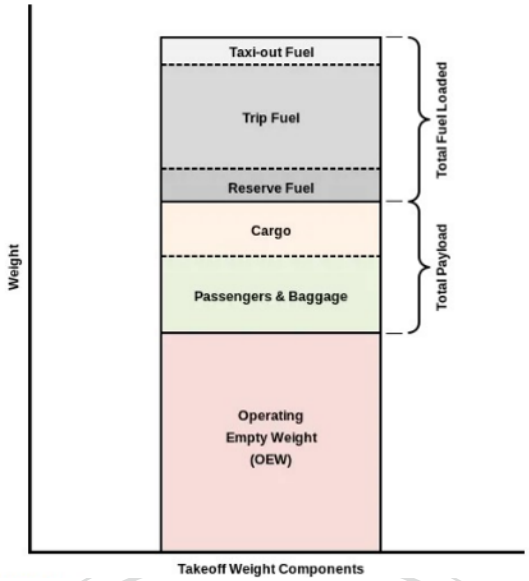

Figure 4-1 Flight weight composition

4-2 Related Work

This chapter examines past research on calculating or measuring the relationship between fuel and transportation in the transportation field and further examines the calculation of transportation fuel control in the aviation field.

The problem of using operational data for statistical modeling of the cyber-physical system was addressed in a case study of an aircraft engine in Gaussian et al (2017). The data from flight data records are used to model the fuel flow rate. This study provides a fundamental understanding of aircraft fuel control and suggests different variables that affect fuel control.

We also make reference to artificial neural network research in transportation management and fuel consumption. For example, a representative neural network model to aid fuel consumption was developed using data given in an aircraft performance manual (Trani et al., 2004). Using Data Envelopment Analysis (DEA) to investigate whether the high level of cargo operations of the world's major airlines would improve the operational efficiency of mixed passenger/cargo airlines (Hong & Zhang, 2010). Application of Machine Learning for Fuel Consumption Modelling of

DOI:10.6814/NCCU202101328

Trucks by collecting various truck and road characteristics (Perrotta et al., 2017). This type of research helps us to understand the relevant features needed to model the fuel consumption prediction in flight.

4-3 Dataset

This case adopts the flight data set from an Asian airline company. There are four data resources, flight info with 6615 instances and 8 features (flight number, season, period, week, origin, destination, aircraft type, number of passengers per cabin); passenger flight details with 1,251,241 instances and 15 features (Nationality, ticket type, age, zodiac sign, employee, gender, travel area code, sector number, group travel, class of flight, seat type, transfer, membership, boarding); baggage details with 1,057,197 instances and 4 features (Number of bags, weight, weight unit, weight limit); travel area with 1,412,877 instances and 4 features (Travel area serial number, travel subarea serial number, travel departure area, travel arrival area).

4-4 Data Preprocessing

Feature Selection

The selection of the columns for flights can result in too few flights in some categories if there are too many variables in the relevant categories. After considering the number of data, we chose to include seasons and flights (different departure points are recorded as different flights) in the model. According to Rose et al. (2012) analyzed the effect of gender, age and class of passengers on airline selection. Therefore, for the passenger-related fields, we chose to differentiate by gender, age, and class of travel into male adult, female adult, male child, female child, and infant.

The data we used in this case consists of takeoff and landing locations, seasons, aircraft type (wide-body or narrow-body), passenger age and passenger gender. We

DOI:10.6814/NCCU202101328

have 6615 instance and selected 10 features for the training. In this study, the season, the aircraft type, male adult, female adult, male children, female children, infants, Cabin, departure, arrival are the characteristics of training data which are more associated with checked baggage weight (Jiang & Zheng, 2020; Nicolae et al., 2017).

To avoid excessive cross joins of data tables to produce dimensionally large and sparse data sets. In this case study, the focus of the problem is on the total baggage weight of each "flight", so our final merged dataset focuses on the flight and records not only the basic information of the flight but also the composition of the passengers and the class and age of the passengers. We merge passenger flight detail with baggage detail by passenger_ID, and identify travel area information from travel area, and finally merge the data table with flight information by flight_ID. The information of individual flights and the composition of passengers, as well as the total weight of baggage for each flight were sorted out.

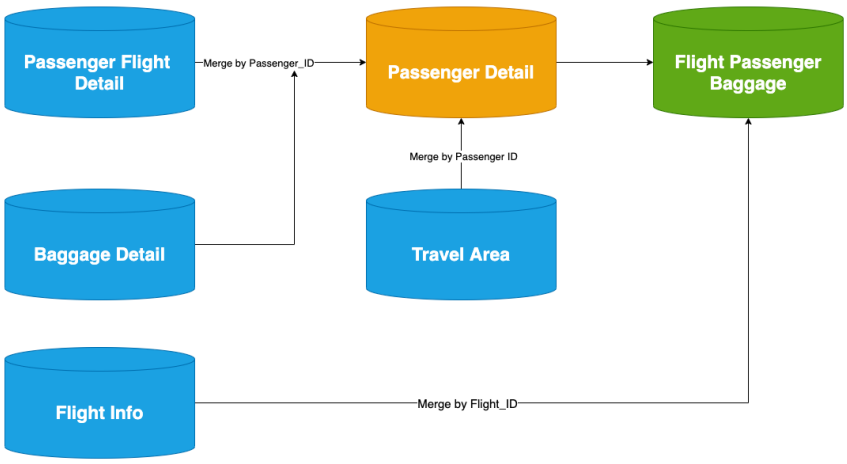

Figure 4-2 Data consolidation

Normalization

Normalization is the scaling of data into a small, specific range. It is often used in certain comparisons and evaluations of metrics to remove the unit constraints of the

DOI:10.6814/NCCU202101328

data and convert them into pure values with no scale, so that metrics of different units or magnitudes can be compared and weighted. In this dataset, the normalization speeds up the gradient descent to find the optimal solution and improves the convergence speed of the model. The data is processed by min max normalization, which scales the data to the [0,1] interval, and the MinMaxScaler suite of scikit-learn is used. Finally, we cut the selected data into 5,292 training data (80% training data) and 1,323 test data (20% test data) by using train_test_split in scikit learn.

4-5 Model Selection and Training

In this section, we introduce the machine learning methods we use. We choose to use decision trees and regression trees because they can predict numerical data. We also use neural networks as one of the prediction models, and we find the best and most suitable model for this case among the three methods.

Random Forest. The basic principle of Random Forest is to combine multiple CART trees, which are decision trees using the GINI algorithm, with randomly assigned training data to significantly improve the final computation. The use of random forests is mainly to deal with classification and regression problems and to improve the accuracy without increasing the computational power. Advantages include handling missing values and filling them, maintaining high accuracy even with large amounts of missing data, efficiently handling small amounts of data, being useful for data mining, detecting outliers, and data visualization. The drawback is that it is overfitting in some classification and regression problems with large noise. In short, Random Forest can be considered as an extension of Decision Tree.

XGBoost. XGBoost is a gradient boosting decision tree that can be used for classification and regression problems. Gradient boosting corrects the residuals of all previous weak learners by adding new weak learners, and finally adding multiple

DOI:10.6814/NCCU202101328

learners together for the final prediction. The performance is more accurate than single machine. It is called gradient due to the fact that gradient descent algorithm is used to minimize the loss while adding new models.

Neural Network. Artificial Neural Network, in the field of machine learning and cognitive science, is a mathematical or computational model that mimics the structure and function of biological neural networks and is used to estimate or approximate functions. A neural network is a large number of artificial neurons linked together to perform computations.

Then, we choose Root-Mean-Square Error (RMSE) as the criterion for evaluating the model, which is the square root of the ratio of the square of the deviation of the observed value from the true value to the number of observations. RMSE is very sensitive to very large or very small errors in a set of measurements. We also use R square as an index to evaluate the accuracy. R square is metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. Therefore, RMSE and R square can reflect the precision of the measurement very well.

Table 4-1 Model RMSE and R square

	MAE	RMSE	R Square
Random Forest	.035	.05126	.915
XGBoost	.034	.04761	.929
Neural Network	.0599	.0599	.816

4-6 Explanation Framework Implementation

Based on the previous framework design, we develop explainable business-supported visualization systems to enhance end-user trust in machine learning. We aim

DOI:10.6814/NCCU202101328

to explore how real users interact with explanations generated by models built on real datasets to understand any subtle differences in AI decisions. We use airline datasets to train multi-labeled gradient boosted trees (XGBoost). The explanation framework proposed in this study is used for explanation. We use advanced XAI tools, and the following introduce the different stages of machine learning for interpretation.

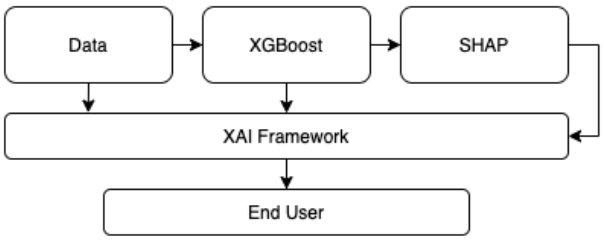

Figure 4-3 The relationship between "data processing and prediction" and the explanatory framework.

What. The input and output data are explained. First, in the input, we present the size, type, and dimension of the training data, and present them as pie charts, numbers, and text. A total of 6615 entities of numerical type and classification are used in this dataset. In addition to the data appearance, we also provide explanations for the selection of features. We present the correlation of features by correlation graph in Figure 4-4 to demonstrate the relevance of the selected elements to the prediction results. Then, in the interpretation of the output results, because of the uncertainty of commercial numerical data, we present them in the form of ranges.

How. According to the proposed framework, the interpretation of the full domain informs the adjustment of the parameters of the user model. In this case, the passenger-related fields are highly correlated with each other, but the explanatory power is slightly different. Therefore, we use Principal Component Analysis (PCA) to linearly combine the passenger-related variables to achieve dimensionality reduction, and implement

DOI:10.6814/NCCU202101328

SciKit Learn, a powerful Python library containing many classical machine learning algorithms and datasets. This stage also provides the model performance scores, in this case the predicted result s are numerical data (luggage weight), so we use RMSE and MSE to determine the numerical scores produced. Finally, our RMSE in random forest is 0.052 with an accuracy of 91.5%, in XGBoost RMSE is 0.051 with an accuracy of 91.7%, and in neural network MAE comes to 0.096 which is much higher than XGBoost and random forest.

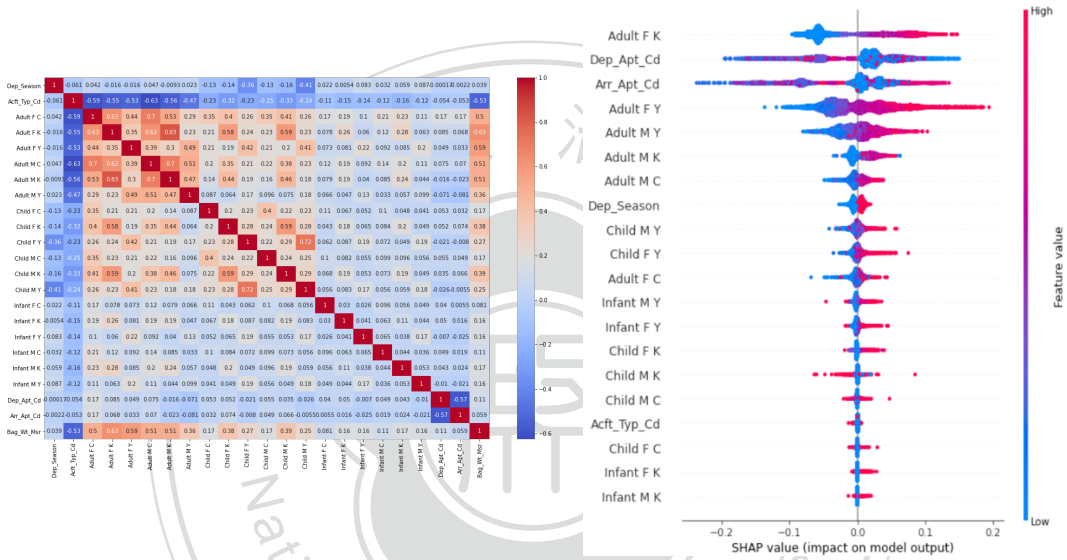

Figure 4-4 Correlation of selected feature. SHAP summary plot of a 23 feature XGBoost prediction model on baggage weight prediction.

Why. At this stage, we use the state-of-the-art XAI explanatory tool SHAP for localized interpretation based on the interpretation framework and comparison table proposed in Chapter 3. SHAP has a solid theoretical foundation in game theory to come up with an explanation. It introduces a way to connect LIME and Shapley values so that we can have a good Kernel Explainer method to explain any model (Lundberg & Lee, 2017). We choose SHAP because it can be interpreted at the individual level and can also aggregate results to obtain local and global interpretations. Although LIME has the same post hoc interpretation feature, LIME generates interpretation based on

DOI:10.6814/NCCU202101328

perturbation samples and does not have a native tree structure interpretation method. In this case we use a tree model for prediction with mainly numerical data, and SHAP has a fast implementation of a tree-based model called Tree Explainer. Fast computation helps SHAP become popular and well used (Lundberg & Lee, 2017). This property is compatible with the model we use (XGBoost) and is one of the main reasons why we chose this explanatory algorithm. The tornado plot in Figure 4-4 is used for visual presentation, providing end-users with an understanding of which characteristics are affected by the prediction results. In this case, our model not only considers flight origins as an important influence but also considers adult female passengers in the premium economy as an important influence. We combine the interpretation results of SHAP with text narratives to deliver more easy-to-understand information to users. This type of information is revealed to help business units add empirical judgment to their baggage weight calculations, effectively increasing their confidence in the prediction results.

DOI:10.6814/NCCU202101328

Chapter 5 Experiment

A between-subject experiment was conducted to verify that this explanatory framework is effective in enhancing users' trust in machine learning. The experiment was designed to simulate a business unit decision process in the business case.

5-1 Task and material

We design a baggage forecasting task based on a simulated airline business requirement scenario, where participants are asked to determine the accuracy of different forecasting methods based on some flight statistics and passenger data. The data used for this task is the aforementioned dataset, which contains 6615 entities, each described by 23 attributes. The total weight of checked baggage for these flights has been recorded as a whole value that are used to evaluate the basis of the participants' acceptance of the prediction. We intend to create a setup that is close to a real-world AI-assisted decision-making scenario. First of all, we provide the original calculation method of the airline as a reference basis. According to the original airline workflow, the calculation is (number of boarding passengers * average weight of checked baggage over the past passengers). The interface presents the raw data including the number of passengers on the flight, and flight information such as route, season, and aircraft type in Figure 5-1. The calculation results are also presented on the right side of the screen.

Machine Learning Model

We use 23 attributes as training features and as prediction features displayed to the participating profiles in the experiment. The model room is randomly assigned for training based on 80% of the original data set, and the remaining 20% of the test data is used for prediction in the experiment. The interface presents the prediction result values and basic information in Figure 5-1.

DOI:10.6814/NCCU202101328

Machine Learning with XAI Framework

We take the prediction results of the previous stage of machine learning predictions and add an explanatory framework in Figure 5-1. We present the information of different stages of machine learning to the end-user. The "What" stage presents the basic information, the data input, the selected features, and the data association. The "How" stage includes the accuracy of the prediction. Finally, in the "Why" stage, we present the contribution of each feature to the results through SHAP Value, and describe in words which factors mainly affect the prediction results.

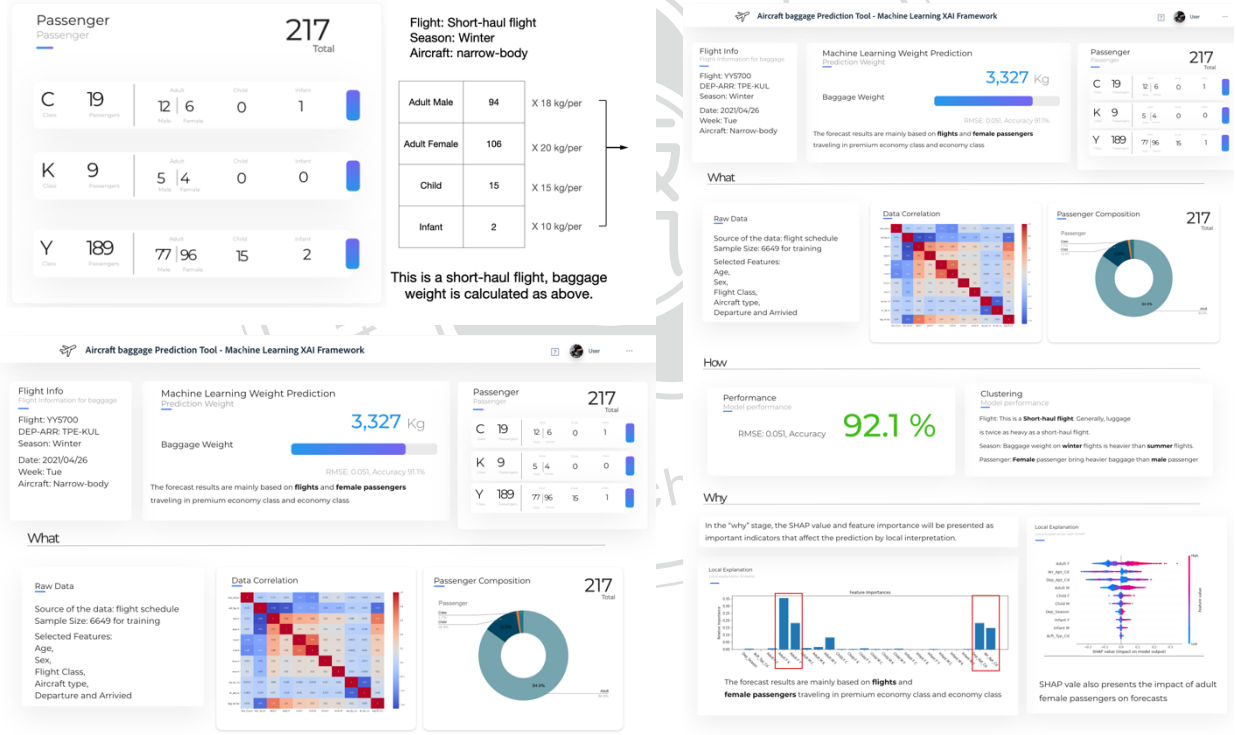

Figure 5-1 On the upper left, the traditional calculation method and presentation. Bottom left, machine learning prediction and presentation. On the right, machine learning prediction combined with explanatory framework.

We conduct a between-subject experiment. This experiment aims to evaluate the effectiveness of the explanation framework through two tests: "ML prediction" and

DOI:10.6814/NCCU202101328

"ML prediction with explanation framework".

First, the experiment evaluates the increase of users' trust in the output of the system through numerical comparison. After seeing the predictions and explanations of the model, the experiment allows participants to manually enter numbers so that users can make more intuitive judgments, and record this value as a reference indicator of whether the user's conversion rate has increased. Participants first be informed of the results of the airline's traditional calculation methods and then are asked to give a fixed value. Next, the user refers to the explanation provided by the different conditions and give a value of his own decision. Each participant completed the prediction of 3 cases. After the experiment is completed, we ask users to do a trust evaluation question, collect whether the trust has been improved, and compare the contribution of different elements in the interpretation framework to the trust improvement trial user (Madsen & Gregor, 2000).

5-2 Participant and experiment procedure

To assess the impact of the explanatory framework on human trust, we conducted experiments and trust surveys to deepen our understanding of the research questions. For this study, 524 participants were recruited on Amazon Mechanical Turk to conduct a quantitative study. In selecting participants, we restricted participants to those from English-speaking countries such as the United States, United Kingdom, Canada, India, and Australia, etc. This is to ensure that our results are conducted in an English-speaking behavioral model. We also limit participants to those who have completed a certain number of HITs (Completed more than 10 hits and have not been rejected) in Mturk and have a "Master qualification" (Average approve rate > 90%) status to participate in our prediction tasks. The experiment continues to collect responses for one week. And our questionnaire was conducted using SurveyCake and deployed to

DOI:10.6814/NCCU202101328

Mturk. On the Mturk page, we designed a link to our questionnaire and briefly explained the questionnaire case. The link was set as a fair random condition, and participants were assigned to one of the two conditions unknowingly.

All participants conducted the experiment independently. Each participant received a flat payment of $0.10. The experiment was approved by specific questions and the correct completion code.

The overview of our proposed experimental process:

Step 1: Participant recruitment. We recruited 524 participants in Amazon Mechanical Turk. We removed technical questions or responses with short response times (response time < 120 seconds). We also designed a manipulation check question in the survey where we asked participants to answer the total number of passengers on the flight that could be found directly on the page, and eventually 13% of the results were removed from the record. All analysis were conducted on the remaining 453 participants. Participants were randomly assigned to examine either a model that explains the predictions through the XAI framework or a model that explains the predictions in general. All participants saw the same set of flights (i.e., the same eigenvalues). The models were constrained to make the same predictions for these flight baggage weights, so that participants saw the same model predictions regardless of the experimental condition to which they were assigned. In addition, the accuracy of the models is almost identical. Thus, the variation between experimental conditions is simply a difference in interpretation. This is a key feature of our experimental design that allows us to run tightly controlled experiments.

Step 2: Preliminary survey. In this stage, our survey collected personal information of users, including age, gender, education level, and understanding of machine learning.

DOI:10.6814/NCCU202101328

Step 3: Experiment. First, the participants were shown a detailed description, including the corresponding plain English description under the condition of a clear model. We provide basic judgment, such as the weight of the baggage may be affected by season, passenger gender, etc. And a linear calculation method has been provided to construct a basic knowledge of airline baggage calculation for the test subjects. Participants were asked to predict the baggage weight of the flights predicted by the model and were asked how confident they were in the prediction. The confidence score was assessed on a 7-point scale. Next, the model predicted weight of the baggage of the flight is displayed, and the participants determine the final baggage weight after seeing the model prediction. At the end of the test, users were asked to fill in their confidence in the model's predictions and their confidence in their own correct predictions. Confidence scores were also assessed on a seven-point scale. In order to ensure that participants understand these instructions, each participant conduct a training case before conducting the experiment, followed by three test cases at the end of the training.

- *User training.* Since the participants may not be familiar with this task, we introduce the case scenario to the participants and allow them to make decisions from the perspective of an aviation practitioner in Figure 5-2. After explaining the usage scenario, the user was asked to perform a test case to ensure that the user was able to perform the experiment correctly. At the end of the training phase, the actual weight of the flight is displayed.
- *Testing.* After completing the training, participants were asked to complete three weight prediction cases. In the testing phase, participants were presented with new flight information. All participants saw the same set of flight information, as random selection may generate additional noise and reduce the efficacy of the experiment, making it more difficult to detect differences between experimental conditions.

DOI:10.6814/NCCU202101328

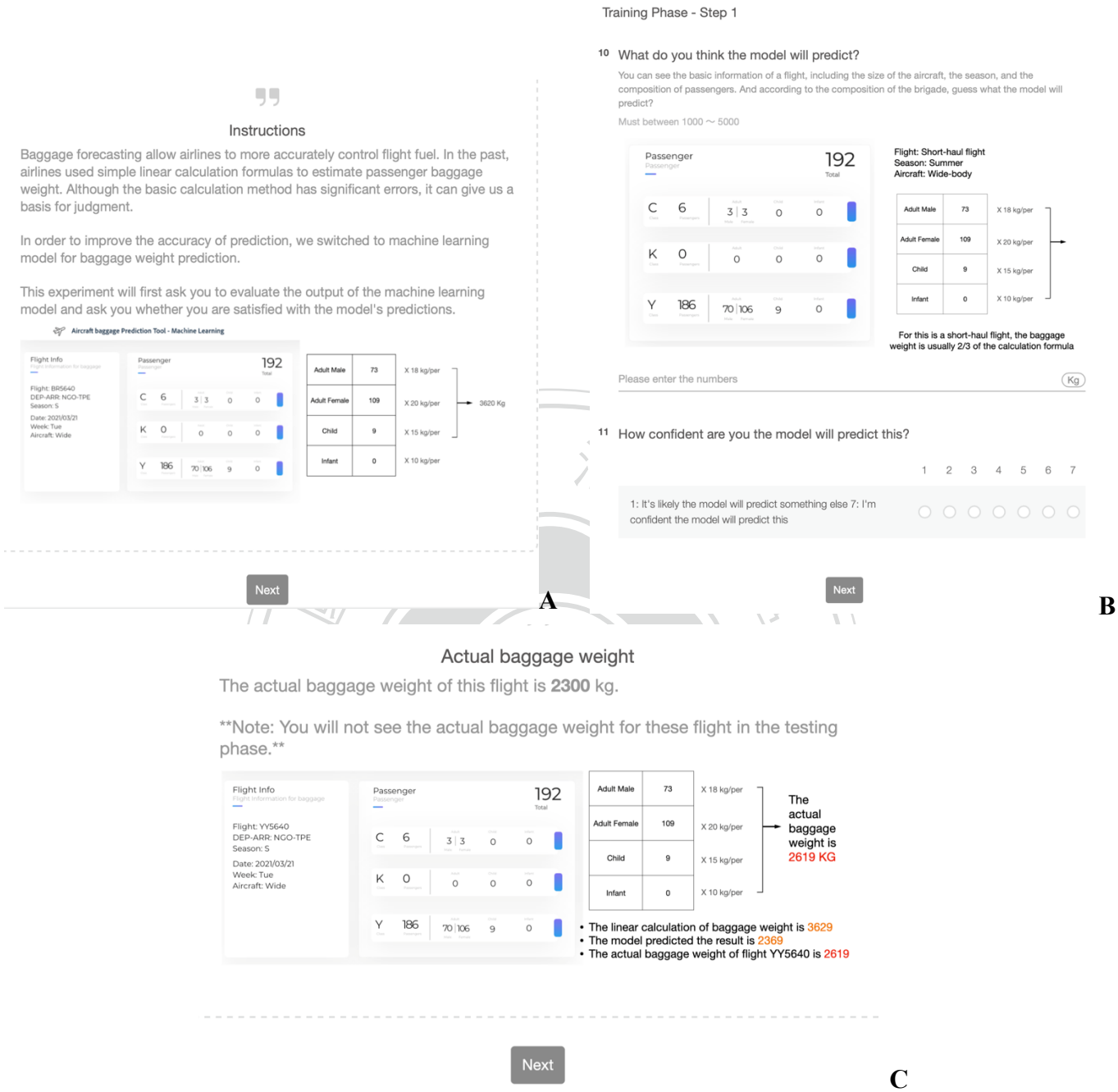

Figure 5-2 Experimental training description.

Step 4: Posttest. Following the experiment, users were presented with a survey of five 5-point Likert scale questions that they could choose to complete in Figure 5-3 Given previous evidence that AI explanations affect understanding and trust, these questions assessed the key dimensions of trust that are widely used in trust issues with respect to system understanding, the capability to assess the system, whether there is faith in the

DOI:10.6814/NCCU202101328

system, reliability ,and the benevolence of the system.

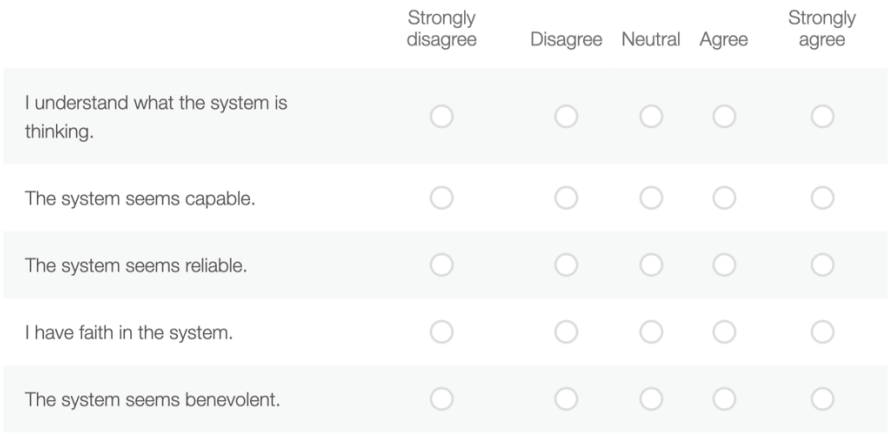

Figure 5-3 A 5-points Likert scale questions after experiment

5-3 Measurement

This study focuses on the XAI framework designed for end-users to enhance their trust in the machine learning results. To evaluate the effectiveness of the interpretive framework and to answer our research questions, we conducted a business case of airlines calculating baggage weight. Then we apply the explanatory framework from Chapter 3. We measured the participants' predicted weight values and analyzed the users' confidence in the explanatory framework. We considered this element because confidence and trust are important factors for users to accept and use the process output (Pieters, 2011). Finally, the validity of the framework is evaluated by means of a questionnaire.

Totally, we collected the deviations of participants' predictions from the model and asked the participants their confidence in the prediction results after each prediction. Then, we conducted a 5-point Likert scale based on the human-computer trust calculation proposed by Madsen & Gregor (Madsen & Gregor, 2000). Participants were asked to rate the understanding, confidence, reliability, strength of the model, and

DOI:10.6814/NCCU202101328

benevolence of the model in Figure 5-3.

Understandability means that a human supervisor or observer can form a mental model and predict the future behavior of the system.

Capability of the system means that the system is considered to be able to perform the task accurately and correctly based on the input information.

Reliability of the system, in the usual sense of repetitive, consistent functionality.

Faith means that the user has confidence in the future capabilities of the system, even when it has not been tried.

Benevolent means that the emergence of machines is just to do their own work, help humans, and be able to provide services that are approachable.

DOI:10.6814/NCCU202101328

Chapter 6 Experiment Result

An experiment of between subject analysis was conducted to test the hypotheses. We conducted a between-subject of two conditions. In total, 453 participants were conducted in our experiment of which 279 were male (61%), 71 were younger than 25 years old, 213 were between 26 and 35 years old, and 167 were older than 36 years old. There are 386 people who have a basic understanding of machine learning.

Table 6-1 Demographic characteristics of participant

Category		Number of respondents	(%)
Gender			
	Male	279	61.8
	Female	172	37.9
	Refuse to answer	2	0.4
Age			
	Less than 20	3	0.6
	21-25	68	15.0
	26-30	107	23.6
	31-35	106	23.3
	36-40	53	11.6
	41-50	71	15.6
	More than 50	43	9.4
Education Level			
	High School	44	9.7
	Undergraduate	106	23.1
	Graduate	303	66.8
ML Level			
	Commonly	66	14.6
	Familiarly	136	30
	Masterly	250	55.3
Country			
	India	202	44.5
	United States	168	37
	Canada	22	4.8
	Other	61	13.4

DOI:10.6814/NCCU202101328

Our proposed explanatory framework can be verified by analyzing this online questionnaire experiment. We also wanted to know the participants' understanding of this visualization framework, so we included open answers at the end of the questionnaire to collect participants' opinions and different views on this framework. After conducting the experiment, we compared the behavior of the participants under different conditions. The experiment is divided into basic model and XAI architecture model. To measure the participants' trust and confidence in the explanatory framework, we analyze it in three main sections. First, in deviation, we analyze the participants' weight prediction for each case. Then, we analyze the confidence scores of users in the model and the differences in confidence scores. Finally, we analyze the post-test of the experiment. We found significant differences in the responses of participants under different conditions. Our findings are as follows:

Deviation.

We define each participant's prediction bias as $|w - wm|$, which is the deviation of the participant's prediction of the final outcome after receiving the explanation, i.e., the absolute difference between the participant's predicted weight w and the number wm given by the model. Such a difference value can be used to assess whether users are more willing to accept the predictions given by the model according to Poursabzi-Sangdeh et al. (Poursabzi-Sangdeh et al., 2021) research. To calculate the predicted weight difference, we used independent sample T-test for analysis. Figure 6-1 shows the final predicted weight w made by the participants of different conditions in the three cases, after seeing the predictions given by the model. Although participants that saw the XAI interpretation gave closer values, we did not find significant differences between predictions from the participants in the traditional condition and those in the detailed XAI mode condition ($\mu_{basic} = 502.64$ *vs.* $\mu_{xai} = 418.97$, $p = 0.075$). The result may indicate that the subjects preferred to fill in the final predicted weight given by the

DOI:10.6814/NCCU202101328

model in both explanatory frameworks.

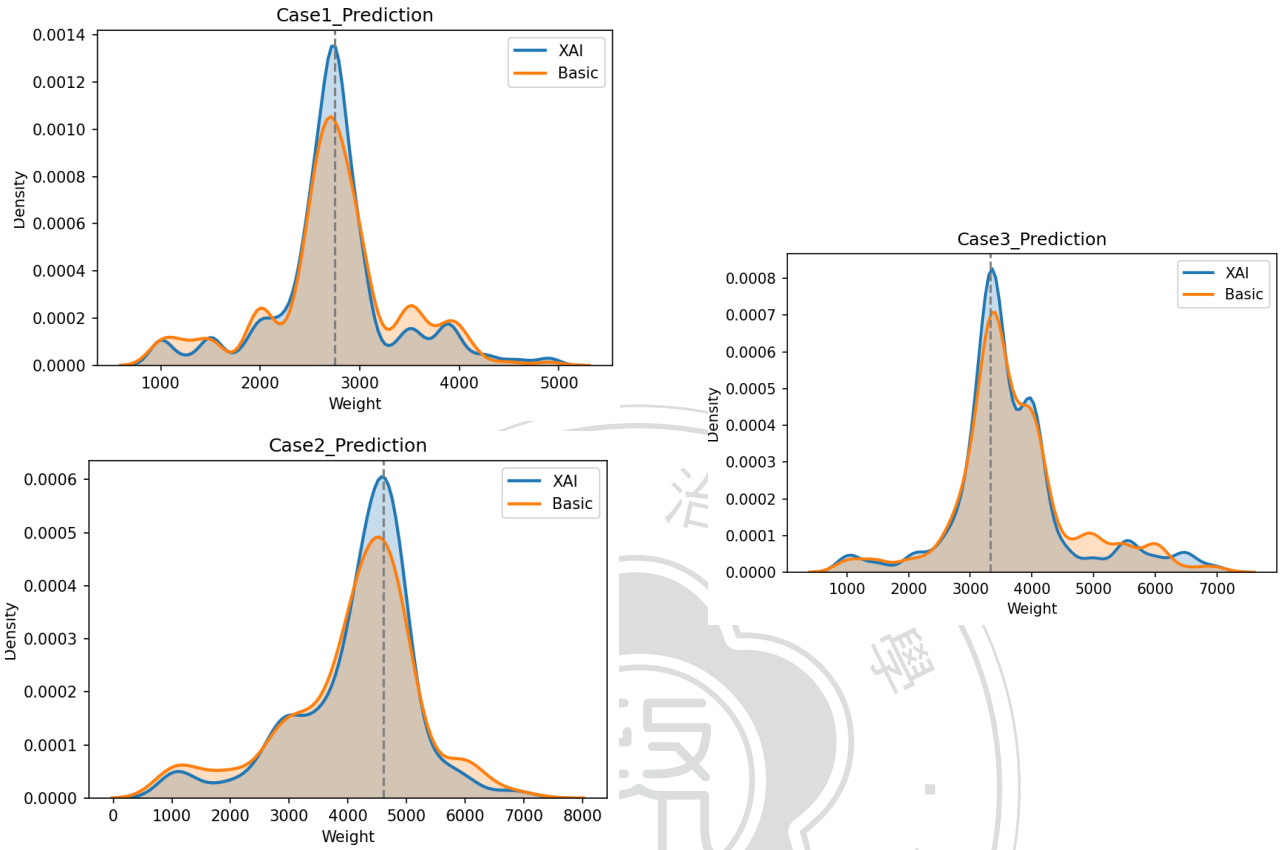

Figure 6-1 Density map of participants' prediction of weight in three cases.

Table 6-2 chi-square tests of deviation in each case.

Case	Conditions	Not Follow	Follow	*P*
Case1	Basic	115	96	.051
	XAI	108	133	
Case2	Basic	103	108	.397
	XAI	108	133	
Case3	Basic	97	114	.221
	XAI	125	116	

To measure whether users follow the predicted values given by the model, we refer to the study by Zhang to determine whether users follow the predicted values of the

DOI:10.6814/NCCU202101328

model in a categorical way (Zhang et al., 2020). We took the median weight response value of each case as an indicator of user acceptance and conducted chi-square tests. (Manikandan, 2011). The results show in Table 6-2 that there is no significant difference in the users' opinion on whether to follow the model results under different explanatory models.

Prediction Confidence

After the participants saw the model results and explanations, we collected the participants to understand their confidence in the correctness of each prediction. The data analysis was performed using generalized estimating equations (GEE) supported by SPSS software. GEE can appropriately specify repeated measures for multiple types of information, allowing us to define distributions and link functions to linearly model different types of outcome variables.

To calculate the association of participants' confidence in the model under different explanatory frameworks, the participants' degree of machine learning may affect their confidence in the explanatory framework, so we also considered the participants' machine learning knowledge. In this calculation, the dependent variable is the participant's confidence score in the explanatory framework, within-subject is the classification by case, and the framework and the ML degree of the participant are the factors.

Table 6-3 GEE model effect

Source	Wald Chi-Square	df	*p*
Framework	19.850	1	<.001
ML Level	87.874	2	<.001
Framework * ML Level	2.862	2	.239

DOI:10.6814/NCCU202101328

Table 6-4 Estimated Marginal Means of Frameworks

Framework	Mean	Std.
Basic	4.53	.068
XAI	4.91	.055

Table 6-5 GEE parameter estimates

Parameter	Wald Chi-Square	Std	Exp(B)	*p*
Basic	4.613	.0933	.808	.032
XAI	-	-	1	-
ML Level 1 (Commonly)	19.482	.1424	.533	<.001
ML Level 2 (Familiarly)	27.118	.1055	.577	<.001
ML Level 3 (Masterly)	-	-	1	-

The result shows in Table 6-3. The model exists significant effects on both framework and machine learning level factor ($p < .001$). There was no significant difference in the interaction between the two factors ($p = .239$). Table 6-4 showed estimated marginal means of two explanation framework. The results show that XAI framework has a significantly higher confidence score compared to basic framework ($p < .001$). We conducted a within-subject experiment to investigate whether different machine learning levels of participants and explanation frameworks would affect users to evaluate confidence. We found differences in the explanatory framework and in the machine learning level of participants. Table 6-5 shows the parameter estimates for the two factors. Participants assigned to Basic had a 19.2 percentage point lower confidence score compared to those assigned to XAI. In terms of participant machine learning level, ML level 3 participants generally had higher confidence in the model than level 1 and level 2 participants. This result shows that not only the explanatory framework of the model has an impact on participants' confidence, but also the participants' own level of machine learning affects whether they have confidence in the model.

DOI:10.6814/NCCU202101328

Confident Deviation.

We used the change of confidence scores in each case to calculate the increase or decrease in user confidence. The first confidence score measures the confidence that the participant got the answer right before seeing the model prediction, and the second confidence score is the confidence that a participant gave the correct answer after seeing the model prediction. In order to confirm the change in user confidence scores, we performed a paired sample T-test for each of the two conditions. The paired-sample t-test was used to compare the mean difference between the two "dependent samples". The results show that in Table 6-6, participants assigned to the XAI interpretation showed a significant increase in confidence scores, while participants assigned to the Basic interpretation showed a significant decrease in confidence in the correctness of the results.

Table 6-6 Confident deviation.

Conditions	Pair	Mean	N	Std. Deviation
Basic	Confident (C)	4.95	633	1.461
	Confident After (CA)	5.73	633	1.484
	CA-C	-.216	*Significant at p < .001*	
XAI	Confident (C)	4.95	723	1.275
	Confident After (CA)	5.16	723	1.407
	CA-C	.203	*Significant at p < .001*	

Table 6-7 Estimated Marginal Means of two framework.

Framework	Mean	Std. Error	*P*
Basic	-.193	.060	< .001
XAI	.194	.063	< .001

DOI:10.6814/NCCU202101328

Then, we evaluated the difference in confidence scores between groups by two-way analysis of variance (ANOVA). ANOVA is used for analysis of variance with only one independent variable, and to compare whether there is a significant difference between groups. The results show that in Table 6-7, there is a significant difference between the two groups, with an increase in confidence in the XAI framework ($p < .001$) and a decrease in confidence in the Basic explanation ($p < .001$). Figure 6-2 shows that ML level 3 has the largest and most extreme change in participant confidence scores, but in fact, there is no significant difference in the interaction between the explanation framework and participant ML level ($p = .629$), as shown in Table 6-8. Additional findings in Table 6-9 we found that ML level 3 participants differed significantly from ML level 2 participants under the XAI explanation (Mean Difference = .243, $p=.035$). Table 6-10 presents the differences between participants with different levels of machine learning in the two conditions, and the results show that there are significant differences between ML level 3 participants in the two conditions.

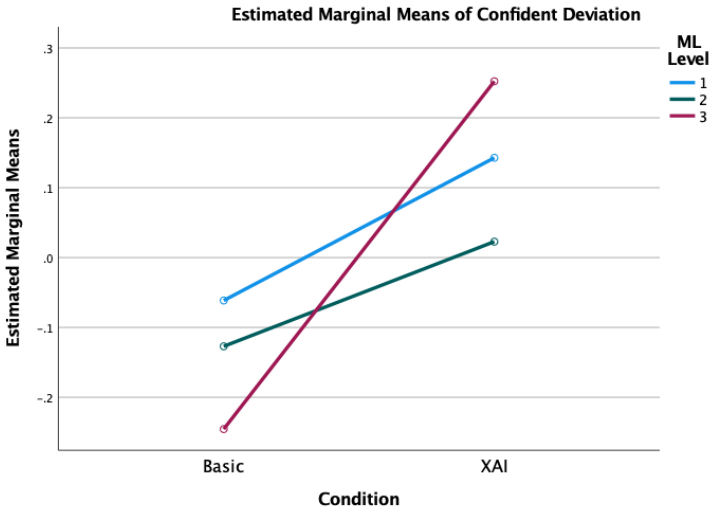

Figure 6-2 Estimated Marginal Means of Confident Deviation.

DOI:10.6814/NCCU202101328

Table 6-8 Estimated Marginal Means of framework and ML level interaction.

Framework	ML Level	Mean	Std.
Basic	1	-.140	.130
	2	-.159	.101
	3	-.279	.076
XAI	1	.202	.151
	2	.068	.094
	3	.312	.068

Table 6-9 Pairwise Comparisons of Confident Deviation by framework

Framework	ML Level(I)	ML Level(J)	Mean Difference(I-J)	*P*
Basic	3	1	-.138	.35
		2	-.120	.34
XAI	3	1	.110	.51
		2	.243*	.035

Table 6-10 Pairwise Comparisons of Confident Deviation by Machine learning level

ML Level	Framework(I)	Framework(J)	Mean Difference(I-J)	*P*
1	XAI	Basic	.343	.086
2	XAI	Basic	.227	.099
3	XAI	Basic	.597*	>.001

Time. Time can be used as an indicator of how much time users spend on the explanation page to process messages and complete tasks. We collect the time participants spend on the entire questionnaire task. Compared to the basic condition, users who were given the XAI framework explanation spent more time. After running the significance test in Table 6-11, the results show that the XAI explanation makes users spend more time viewing the content on the page (μ_{basic}=*825.20 sec. vs.* μ_{xai}=*911.58 sec., p = .024*). We also conducted a two-way ANOVA to assess whether there was a difference in the time spent on the test between participants of different ML

DOI:10.6814/NCCU202101328

levels. The results showed that the ML level did not make a significant difference in the time difference ($p = .481$).

Table 6-11 Statistical properties from two interpretation modes.

	Basic	**XAI**
Average working time (sec.)	825.20	911.58*
Std. Deviation	471.29	390.99
Number of turkers	211	241

**Significant at p < .05*

Table 6-12 Mann-Whitney U test of the post-test questionnaire

	Conditions	**N**	**Mean Rank**	**Sum of Ranks**
Understand	Basic	211	229.85	48498
	XAI	241	223.57	54880
Capability*	Basic	211	214.09	45174
	XAI	241	237.36	57204
Reliability*	Basic	211	214.18	45193
	XAI	241	237.28	57185
Faith**	Basic	211	203.88	43018.5
	XAI	241	246.30	59359.5
Benevolent	Basic	211	221.76	46791
	XAI	241	230.65	55587

** Significant at p < .05, ** Significant at p < .001*

Posttest. After the experiment is over, users are asked to complete a 5-point Likert scale question survey. The distribution of groups in the posttest questionnaire of this study is unknown and may be non-constant. Therefore, we used the nonparametric statistics Mann-Whitney U test to conduct the analysis. The Mann-Whitney U test is used instead of the t-test to determine whether the means of the two independent normative

DOI:10.6814/NCCU202101328

distributions are equal, and the U-test does not require the assumption of normative distribution and is suitable when the distribution of the clusters is unknown and not normative. In view of the previous evidence that AI interpretation affect understanding and trust, these questions assess the key dimensions of trust in Table 6-12. The results are shown as follow.

- *Understand.* After seeing the model's prediction of weight, the participants of both explanatory methods considered that they could understand the operation of the system and that the system gave them sufficient explanation, and there was no significant difference in the two conditions ($U=24719, p = .587$).
- *Capability*. In terms of system capability, if users see the explanation of the XAI architecture, they give a higher rating to the system capability ($U=22808, p < .05$).
- *Reliability*. The reliability of the system is a repetitive and consistent function in the usual sense. Users who accept XAI interpretation under this questionnaire believe that the system provides more reliable services ($U = 22827, p < .05$).
- *Faith*. Faith means that users have confidence in the future functions of the system, even if it has not been tried. In this test, participants who use XAI interpretation have a high level of confidence in the system ($U = 20652.5, p < .001$).
- *Benevolent.* The benevolent machines were designed to do their job, to help people and to be approachable. In both cases, participants did not see a significant difference in the affinity of the system ($U = 24425, p = .449$).

The results show that there is no significant difference between the two explanatory frameworks in terms of understanding and benevolent. In terms of system capability, reliability, and faith, the participants of XAI interpretation gave higher ratings and showed significant differences.

DOI:10.6814/NCCU202101328

Chapter 7 Discussion

The main objective of this study is to provide end-users of machine learning in the business domain with a systematic understanding of the results of AI and to increase their trust in it, thereby increasing the practical application of AI technology in the business domain. The results corresponding to the research questions are presented in order to achieve the research objectives.

7-1 General Discussion

For research question 1, "*Does the use of the XAI framework in the business field allow end-users to follow the predictions of machine learning results?*", previous studies have shown that case-specific confidence information can improve trust calibration in AI-assisted decision-making scenarios and allow users to accept the model's predictions (Zhang et al., 2020). To measure the conversion rate of users after seeing the model, we asked them to freely fill in the values they entered, unlike previous studies that used purely binary judgments, we asked users to fill in numbers. Some findings were obtained in our experimental results. Compared to the basic condition, users of the XAI framework did not change the final prediction significantly, although the weight of the final decision was more concentrated as can be seen in Figure 6-1. This result is consistent with the one presented by Poursabzi-Sangdeh et al. (Poursabzi-Sangdeh et al., 2021). They conducted numerical predictions of housing prices and found no significant improvement in the extent to which participants followed the predictions of the clear model with fewer features compared to the predictions of the black box model with more features. Although there is a tendency for the XAI condition to be concentrated in the weight distribution graph, it is not significant in the statistical results. We speculate that the reason for this are as follows: the behavioral performance of the users basically uses the numbers calculated by the model, possibly because of the lack of prior knowledge and reference base.

DOI:10.6814/NCCU202101328

Research question 2, "*Does the application of XAI framework in the business field increase the confidence of users in adopting machine learning results?*", confidence is the self-assurance of the safety or security of the system without knowing the risks or considering alternatives. We considered this element because confidence and trust are important factors for users' willingness to accept and use the process output (Pieters, 2011). We measured user confidence in two ways. First, we use a confidence score that asks users about the model. Second, we use the change in confidence score to measure whether users' confidence in the model increases or decreases.

The results can prove that the participants in XAI explanatory framework have higher evaluation of the model and are more confident. In the results of confidence scores, we can see that user of the XAI framework generally have more confidence in the model. The XAI framework also shows a significant increase in confidence in the confidence change metric. An additional finding is that users with better knowledge of machine learning have significantly higher confidence deviation scores. These results are similar to the design recommendations made by Le Bras et al. (2018). In this paper on explaining models through Data driven, it is suggested that designers can present explanations through visualization because they find that explanations specific to the underlying data can increase participants' perception of the robustness of the process and understanding of the causes, as well as increase participants' competence and confidence in the process of model operation. As in our visualization approach proposed in the explanatory framework, we use not only graphics to convey information, but also text to explain each stage of the framework.

In the participant section, although we provided end-user-based content, participants who were not in the relevant field (aviation knowledge) and were less familiar with machine learning took more time to understand the explanations. Their unfamiliarity with the explanations and related cases resulted in a lesser change in

DOI:10.6814/NCCU202101328

confidence scores than the more advanced participants. This result is similar to Poursabzi-Sangdeh (Poursabzi-Sangdeh et al., 2021) and Zhang et al. (Zhang et al., 2020) in those participant who are not experts in a particular domain may result in a smaller difference in their confidence scores.

For research question 3, "*Does the application of XAI framework in the business field increase end-users' trust in machine learning results?*", we conducted a posterior analysis on users' trust enhancement. We evaluated trust in five dimensions based on the human-computer interaction trust model proposed by Mayer et al. (1995). It was found that there was no significant difference in the understanding of the system between the users of the two explanations in the Understand category. This result can prove that the basic model, although it does not provide a complete explanation, still contains basic information about the type of input data, the composition of flights, the association graph between data, and the accuracy of the model. This information is useful in providing the user with a basic understanding of the model. The XAI framework provides more detailed information than the basic condition, with localized explanatory diagrams, textual descriptions of model training and feature selection. The participants were also able to understand the model operation, so there was no significant difference in the understanding. According to our open-ended question-and-answer session at the end of the questionnaire, participant was asked to fill in how they assessed each metric in the trust score. Several basic condition participants reported that they understood what the system was doing and that users felt the system was not capable. The user responses were consistent with the statistical results of the post-test questionnaire.

Among the three aspects of system capability, reliability, and confidence in the system, the participants assigned to the XAI interpretation gave significantly higher ratings to the system capability, reliability, and confidence compared to the Basic

DOI:10.6814/NCCU202101328

condition. First, in the “what” section, our XAI framework integrates data integration based on business intelligence. Second, in the “how” section we described the system logic in words as well as the system performance, accuracy, etc. The participants' feedback indicated that this was a factor they trusted and considered the system powerful. Last but not least, in the “why” section we combine localized explanations with SHAP, presenting the most important factors in model training with tornado plot and weight lists, and then combine them with textual descriptions to provide explanations to users.

Finally, there is no significant difference between the two conditions in the benevolent orientation. According to Cai et al. (Cai et al., 2019) the effect of the comparative explanation they used on the benevolence of the system was found. Users who saw the comparative explanation perceived the system to be more benevolent. Although the comparative explanation did not increase the perception of the system's capabilities, users may feel that the algorithm is at least making an effort. However, given the relatively small amount of effect, its practical significance may be limited. In this study, because the two explanations provided the same predicted results, for the participants the system provided explanations in different ways in the same situation without any significant difference. As a result, XAI users perceive the system to be more capable and reliable, and it provides XAI users with a higher level of confidence.

Although the participants assigned to the XAI explanation had a high overall trust level, the participants assigned to the XAI model took longer time in the analysis of time. This may be due to the fact that we provide more information in the XAI explanation and the localized explanation of why may take time to understand for the end-user due to unfamiliarity. We designed textual explanations in the system to help users understand the information provided at each explanation stage. This is also consistent with our open-ended Q&A at the end, and user feedback to “*The different*

DOI:10.6814/NCCU202101328

calculations provided by the system enhance confidence in the system, but are slightly more complicated.". The XAI framework explanation provides more information, but also makes the system more complex and costly to learn.

7-2 Limitation and Future Work

One of limitation of our work is that our experiments focus on one type of stakeholder, using one type of model (XGBoost) in a specific domain (airline baggage weight calculation). Future extensions to other types of stakeholders, other types of models (e.g., rule based, deep neural networks), other tasks (e.g., classification), and other domains (e.g., medical diagnosis, credit risk assessment, judicial decisions, and bail) may lead to different results.

Second, for the participants' understanding of the elements, although we made detailed explanations and test cases, and we also used operational checks to ensure that the users performed the experiments correctly. However, we could not be sure that users fully understood each element of the explanatory framework. This may lead to inaccurate judgments and poor results due to the participants' lack of understanding of the explanatory style or unfamiliarity with the cases. In future experiments, we can adjust the number of training sessions, measure the amount of time participants spend on the experiment under different conditions, or add an eye-tracking device to assess whether users are looking at specific areas and making questionnaires for different blocks or interviewing participants to collect their perceptions of the explanatory framework. Although we cannot currently assess users' understanding of individual elements, the overall explanatory framework has been shown in experiments to increase users' confidence in the model and trust in the system. This can bring real value to enterprises when they need to use machine learning models for prediction and classification.

DOI:10.6814/NCCU202101328

Finally, our experiments were run without process measures as dependent variables, which limited our ability to reflect on the cognitive and meaning-making processes that might be at play. As an example, although we measured participants' confidence scores on the model, these scores were based on how they felt about the content of the explanatory framework. We cannot directly infer from these results whether participants understood why the model provided the content. Qualitative experiments such as interviews, thinking out loud, and process tracing designed to understand why people behave the way they do may help to examine the cognitive and meaning-making aspects of interpretability.

DOI:10.6814/NCCU202101328

Chapter 8 Conclusion

Given the widespread and increasing use of machine learning models, it is likely that people will collaborate with the models to make more and more decisions. When this happens, the need for interpretable models is likely to increase as well. Past research on XAI has mostly focused on algorithm-related (Ribeiro et al., 2016) or on the interpretation of image recognition, e.g., salient maps (Bach et al., 2015). Apart from image recognition, most of the research on end-users has focused only on medical applications (Wang et al., 2019) and these algorithms are mainly designed for machine learning developers (Vassiliades et al., 2021). We compile the explanatory algorithms developed in the past and propose an XAI framework for end-users based on business intelligence, applying this framework to business cases to break the dilemma that individual algorithms can solve specific problems. We believe that the integration of explanatory techniques in machine learning applications in the business domain can bring higher value to enterprises.

Our study reviews the existing literature on different XAI technologies and studies that combine human-computer interaction and trust. We integrate different interpretative approaches proposed in the past and propose an AI interpretability framework based on business intelligence to provide explanations to end-users of business domain operations. The framework supports visualization on business data, presenting the feature values and performance of models, and finding the most impactful features using novel local feature importance metrics. The framework organizes the way explanations have been provided to end users in the past, and we aim to help developers build more user-centric, explainable AI-based systems.

Next, to validate the proposed explanatory framework for practical application in the commercial domain, we applied the framework to a real-world commercial data machine learning use case for predicting airline baggage weight, enabling airlines to

DOI:10.6814/NCCU202101328

more accurately calculate fuel consumption and reduce cost per flight, as well as for them to better communicate their models with stakeholders. This research focuses on a business intelligence-based interpretation approach, which emphasizes the integration of input data, machine learning models, and business processes. Enterprises can extend from the original business intelligence and combine this interpretation framework in different stages, for example, the original BI is explained in the “what” stage, and with the explanation of the “how” and “why” stages, there is an opportunity to enhance the reliability of machine learning technology in business applications.

Our experimental results show that participants who use this explanatory framework are more confident in the model predictions and trust the system, and they are more willing to adopt the suggestions provided by the system. Participants perceived the system to be capable, but at the same time more complex and requiring more time to understand. Despite the integrated features and case studies of this XAI interpretation framework, there is still much work ahead to fully understand the machine learning models for end-users in different domains.

In future research, the framework can be extended from different perspectives, for example, from the model developer, the domain expert, and the end-user, and the linkage of different users provides a more comprehensive explanation of the framework for a more compatible application in practice. Studying the use of AI by different users will ensure that AI approaches are implemented and used responsibly in organizations. Moreover, it will collectively facilitate the development of XAI in practical applications and in the business field.

DOI:10.6814/NCCU202101328

Reference

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K. R., Dähne, S., & Kindermans, P. J. (2019). INNvestigate neural networks! *Journal of Machine Learning Research*.

Allen, W. L. (2018). Visual brokerage: Communicating data and research through visualisation. *Public Understanding of Science*, *27*(8), 906–922.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Conference on Human Factors in Computing Systems - Proceedings*.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020a). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020b). Explainable Explainable Artificial Intelligence (XAI): Concepts,

DOI:10.6814/NCCU202101328

taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*.

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *Annals of Applied Statistics*.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable. *IEEE Transactions on Visualization and Computer Graphics*.

Bruls, M., Huizing, K., & van Wijk, J. J. (2000). *Squarified Treemaps*.

Burton, B., Geishecker, L., Schlegel, K., Hostmann, B., Austin, T., Herschel, G., Rayner, N., Sallam, R. L., Richardson, J., Hagerty, J., & Hostmann, B. (2006). Magic Quadrant for Business Intelligence Platforms WHAT YOU NEED TO KNOW. *Gartner Research*, *January*, 1–5. http://www.gartner.com/technology/about/ombudsman/omb_guide2.jsp

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F1476*, 258–262.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. In *Electronics (Switzerland)*.

Cawthon, N., & Moere, A. Vande. (2007). The effect of aesthetic on the usability of data visualization. *Proceedings of the International Conference on Information Visualisation*.

Chati, Y. S., & Balakrishnan, H. (2017). A Gaussian Process Regression approach to model aircraft engine fuel flow rate. *Proceedings - 2017 ACM/IEEE 8th International Conference on Cyber-Physical Systems, ICCPS 2017 (Part of CPS Week)*.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business

DOI:10.6814/NCCU202101328

intelligence technology. In *Communications of the ACM* (Vol. 54, Issue 8).

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, *36*(4).

Clancey, W. J. (1983). The epistemology of a rule-based expert system -a framework for explanation. *Artificial Intelligence*.

Collins, C. R., & Stephenson, K. (2003). A circle packing algorithm. *Computational Geometry: Theory and Applications*, *25*(3).

Davis, B., Glenski, M., Sealy, W., & Arendt, D. (2020). Measure Utility, Gain Trust: Practical Advice for XAI Researchers. *Proceedings - 2020 IEEE Workshop on TRust and EXpertise in Visual Analytics, TREX 2020*, 1–8.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7).

Deng, X., & Chi, L. (2012). Understanding postadoptive behaviors in information systems use: A longitudinal analysis of system use problems in the business intelligence context. *Journal of Management Information Systems*, *29*(3).

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *ACM/IEEE International Conference on Human-Robot Interaction*.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1).

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. *Ml*, 1–13. http://arxiv.org/abs/1702.08608

Dresner, H. (2001). *Business Intelligence in 2002: A Coming of Age - 103282.pdf*. Gartner.

DOI:10.6814/NCCU202101328

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9).

Gillespie, PhD, MA, RN, T. W. (2012). Understanding Waterfall Plots. *Journal of the Advanced Practitioner in Oncology*, *3*(2).

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. *International Conference on Intelligent User Interfaces, Proceedings IUI*.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*.

Gorchels, L. (2000). The Product Manager's Handbook. In *NTC Business Books*.

Groom, V., & Nass, C. (2007). Can robots be teammates? Benchmarks in human-robot teams. *Interaction Studies*.

Henelius, A., Puolamäki, K., Boström, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: Exploring classifiers by randomization. *Data Mining and Knowledge Discovery*.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*.

Hong, S., & Zhang, A. (2010). An efficiency study of airlines and air cargo/passenger divisions: A DEA approach. *World Review of Intermodal Transportation Research*.

Inselberg, A., & Dimsdale, B. (1990). *Parallel coordinates: A tool for visualizing multi-*

DOI:10.6814/NCCU202101328

dimensional geometry.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11).

Jiang, C., & Zheng, S. (2020). Airline baggage fees and airport congestion. *Transportation Research Part C: Emerging Technologies*.

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*.

Kim, Y. S., Walls, L. A., Krafft, P., & Hullman, J. (2019). A Bayesian cognition approach to improve data visualization. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *34th International Conference on Machine Learning, ICML 2017*, *4*, 2976–2987.

Kosara, R. (2016). Presentation-Oriented Visualization Techniques. *IEEE Computer Graphics and Applications*, *36*(1).

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*, 5686–5697.

Langley, P., & Simon, H. A. (1995). Applications of Machine Learning and Rule Induction. *Communications of the ACM*.

Lapuschkin, S., Binder, A., Montavon, G., Muller, K. R., & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

DOI:10.6814/NCCU202101328

Le Bras, P., Robb, D. A., Methven, T. S., Padilla, S., & Chantler, M. J. (2018). Improving user confidence in concept maps: Exploring data driven explanations. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*.

LeBaron, B. (2001). Evolution and time horizons in an agent-based stock market. *Macroeconomic Dynamics*.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Conference on Human Factors in Computing Systems - Proceedings*.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10).

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. *Proceedings of Eleventh Australasian Conference on Information Systems*, 6–8.

Manikandan, S. (2011). Measures of central tendency: Median and mode. In *Journal of Pharmacology and Pharmacotherapeutics* (Vol. 2, Issue 3).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734.

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1).

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding

DOI:10.6814/NCCU202101328

the physical implications of machine learning. *Bulletin of the American Meteorological Society*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence*.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247. https://christophm.github.io/interpretable-ml-book

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Nicolae, M., Arikan, M., Deshpande, V., & Ferguson, M. (2017). Do bags fly free? An empirical analysis of the operational implications of airline baggage fees. In *Management Science*.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3–5), 393–444.

Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M., & Bertini, E. (2014). The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2211–2220.

Panniello, U., Gorgoglione, M., & Tuzhilin, A. (2016). In CARSs we trust: How context-aware recommendations affect customers' trust and other business performance measures of recommender systems. *Information Systems Research*, *27*(1).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*.

Perrotta, F., Parry, T., & Neves, L. C. (2017). Application of machine learning for fuel

DOI:10.6814/NCCU202101328

consumption modelling of trucks. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*.

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1).

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, *38*(1), 99.

Poursabzi-Sangdeh, F., Goldstein, D. G., & Hofman, J. M. (2021). Manipulating and measuring model interpretability. In *Conference on Human Factors in Computing Systems - Proceedings*.

Power, D. J. (2002). Decision Support Systems: Concepts and Resources for Managers. In *Information Systems Management* (Vol. 20, Issue 4).

Putnam, V., & Conati, C. (2019). Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS). *CEUR Workshop Proceedings*.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325–1332.

Rose, J. M., Hensher, D. A., Greene, W. H., & Washington, S. P. (2012). Attribute exclusion strategies in airline choice: Accounting for exogenous information on

DOI:10.6814/NCCU202101328

decision maker processing strategies in models of discrete choice. *Transportmetrica*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. http://arxiv.org/abs/1708.08296

Sautto, J. M. (2014). Decision Support Systems for Business Intelligence, 2nd edition. In *Investigación Operacional* (Vol. 35, Issue 1).

Saxena, R., & Srinivasan, A. (2013). Business intelligence. In *International Series in Operations Research and Management Science*.

Shafiei, F., & Sundaram, D. (2004). Multi-enterprise collaborative enterprise resource planning and decision support systems. *Proceedings of the Hawaii International Conference on System Sciences*, *37*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*.

Swartout, W. R. (1983). XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*.

Touchette, P. E., MacDonald, R. F., & Langer, S. N. (1985). A scatter plot for identifying stimulus control of problem behavior. *Journal of Applied Behavior Analysis*, *18*(4).

Trani, A. A., Wing-Ho, F. C., Schilling, G., Baik, H., & Seshadri, A. (2004). A neural network model to estimate aircraft fuel consumption. *Collection of Technical*

DOI:10.6814/NCCU202101328

Papers - AIAA 4th Aviation Technology, Integration, and Operations Forum, ATIO.

van Wijk, J. J., & van de Wetering, H. (1999). Cushion treemaps: visualization of hierarchical information. *Proceedings of the IEEE Symposium on Information Visualization*.

Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: A survey. In *Knowledge Engineering Review*.

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*.

Vilone, G., & Longo, L. (2020). *Explainable Artificial Intelligence: a Systematic Review*. *Dl*. http://arxiv.org/abs/2006.00093

Wang, D., Yang, Q., Abdul, A., Lim, B. Y., & States, U. (2019). *Designing Theory-Driven User-Centric Explainable AI*. 1–15.

Wang, J., Gou, L., Yang, H., & Shen, H. W. (2018). GANViz: A Visual Analytics Approach to Understand the Adversarial Game. *IEEE Transactions on Visualization and Computer Graphics*, *24*(6).

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. *ACM/IEEE International Conference on Human-Robot Interaction*.

Wong, W. H., Zhang, A., Van Hui, Y., & Leung, L. C. (2009). Optimal baggage-limit policy: Airline passenger and cargo allocation. *Transportation Science*, *43*(3), 355–369.

Xia, M., Asano, Y., Williams, J. J., Qu, H., & Ma, X. (2020). Using Information Visualization to Promote Students' Reflection on "gaming the System" in Online Learning. *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @*

DOI:10.6814/NCCU202101328

Scale, 37–49.

Yagoda, R. E., & Gillan, D. J. (2012). You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. *International Journal of Social Robotics*, *4*(3).

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *International Conference on Intelligent User Interfaces, Proceedings IUI*.

Yu, K., Taib, R., Berkovsky, S., Zhou, J., Conway, D., & Chen, F. (2016). Trust and Reliance based on system accuracy. *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*.

Yur, E., & Vasil, V. (2013). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering*, *2013*, Article ID 969458.

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

DOI:10.6814/NCCU202101328