Explanations derived from inconsistent case bases using authoritativeness

Book Production MANAGER [a,1], Second AUTHOR [b] and Third AUTHOR [b]
[a] *Book Department, IOS Press, The Netherlands*
[b] *Short Affiliation of Second Author and Third Author*

Abstract. *Post hoc* analyses are used to provide interpretable justifications for machine learning predictions of an opaque model. We modify a top-level model that uses case-based argumentation to this end, altering its definition of best precedent to include an expression of authoritativeness. This improves the capacity of the approach to handle inconsistent case bases, which we argue are to be expected in practice. We experiment with a number of expressions of authoritativeness using three different data sets.

1. Introduction

Both machine learning (ML) and rule-based classification approaches involve a trade off between accuracy and transparency, specifically the ability of end-users to understand decisions (class predictions) [1]. Deep neural networks in particular tend to produce predictions with a high degree of accuracy at the cost of transparency due to their technical complexity. However, the perceived complexity may vary according to a person's level of understanding, so even much simpler approaches might be thought of as relatively opaque by some people. Another reason for poor transparency can be proprietary protection of the approach, which can render even a relatively simple approach utterly opaque. Regardless of the underlying reason, the term 'black box' is often used to refer to a particularly opaque approach [1,2]. A poorly interpretable model is more difficult to trust. It is harder to see its shortcomings, including biases and ethical concerns [3]. Explainable Artificial Intelligence (XAI) is aimed at increasing the transparency of otherwise opaque models [3]. In the case of a classification problem in ML, this entails that we can explain why a certain prediction was made by the model in a particular instance.

Methods of explaining ML decisions vary in a number of respects. A distinction can be made between methods that generate local explanations (explaining individual instances) and those that generate global explanations (explaining a whole model). Some methods have access to the learnt model, while others are model agnostic. We use the term 'justifications' for explanations generated without model access to signify that such explanations do not explain exactly how a decision was reached, but instead they explain the assumptions under which the model's decision can be justified. Justifications are thus

[1]Corresponding Author: Book Production Manager, IOS Press, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands; E-mail: bookproduction@iospress.nl.

not intended as separate predictions. This is related to the notion of *post hoc* analysis, which implies that an explanation is produced after the fact [1]. In this paper, we are concerned with local justifications produced by a model-agnostic, post hoc analysis.

One way to justify a classifier's prediction is to show a most similar case to one whose class is being predicted (the focus case). To this end, Prakken & Ratsma [4] draw on AI & law research to propose a top-level model using case-based argumentation (CBA) to explain black-box predictions based on Horty's model of *a fortiori* reasoning [5,6] and inspired by CATO [7], hereafter referred to as 'A Fortiori Case-Based Argumentation' (AF-CBA). AF-CBA produces a human-interpretable justification of the classifier's prediction by treating its training data as a case base (CB) and comparing it to the current focus case, including the prediction made by the model [4]. The underlying a fortiori assumption is that the focus case should have the same outcome as a precedent case if the differences between these cases only serve to add further support for that same outcome. An argument graph is constructed through a grounded argument game consisting of a fixed set of allowed moves. A proponent defends why the focus case should receive the same outcome as a best precedent (a most similar case) and the opponent argues against this. In doing so, they cite examples and counterexamples from the CB. There are distinguishing moves that set cases apart and moves to downplay these differences. When a precedent case has no relevant differences with the focus case, deciding for the focus case is said to be 'forced'. The effectiveness of AF-CBA hinges in part on the distance measure between cases and any feature selection technique used to promote an interpretable argument graph [4].

Supervised ML approaches require that a data set be annotated, i.e. labelled by some other means so that the model can learn to do the same. Annotators produce a labelled data set specifically for the purpose of training a model. However, annotators can and do make the occasional mistake [8]. Furthermore, labels may also be produced by decision makers—people who produce labels as part of their role in some decision process and who may have conflicting opinions. For example, judges do this when they decide on court cases, with their verdict being the label that is stored in a body of jurisprudence. This does not prevent contradictory classifications, as jurisprudence can contain conflicting opinions and interpretations.

Moreover, depending on the origins of the data set, the feature vector itself may be a subset of all relevant details, thereby potentially lacking necessary data to discriminate between seemingly similar cases [9]. These sources of noise make the labelling seem inconsistent, since identical feature vectors might receive conflicting labels. Under the a fortiori assumption, this notion of inconsistency becomes even broader: a case which is at least as good as another yet receives the opposite outcome is a source of inconsistency. For these reasons, CB consistency is generally not a safe assumption in practice.

AF-CBA does not strictly require that the CB be consistent, but inconsistencies are often due to exceptional cases (with a surprising outcome) and these can be problematic for the explanation due to the focus case being forced for both outcomes. In experiments by Prakken & Ratsma [4], significant portions of a case base had to be ignored in order to make them consistent—namely 0.32%, 11.35% and 3.20% for three different inconsistent data sets. This problem is exasperated by feature selection techniques, which would otherwise benefit the simplicity of AF-CBA's justifications. In conclusion, CB consistency forms a problematic constraint for AF-CBA.

In this paper, we present a modification of AF-CBA that takes into account the degree (which we call 'authoritativeness') to which the CB is consistent in regards to the focus case. This measure is used to prevent inconsistent forcing. Furthermore, it is used to modify the selection of best precedents to cite, as it makes intuitive sense to cite cases with the highest authoritativeness. We investigate the desirability of this through experiments with several possible alternatives of quantifying authoritativeness, demonstrating it to have a beneficial effect on the explanations. The rest of this paper is structured as follows. We describe AF-CBA and its background in Section 2. We consider how to address the problem of inconsistency in Section 3. We subsequently experiment with our proposed solution in Section **??**.

2. Case-Based Argumentation

In this section, we present the CBA framework by Prakken & Ratsma (with some differences in notation) for explanations with dimensions [4]) for the purposes of justifying ML predictions as a type of post hoc analysis. As our running example, we make use of the Telco Customer Churn data set [10], which describes the customers of a telecommunications provider and whether or not they churned (switched providers).

Table 1 describes the features/dimensions used. The tendency of a dimension reflects whether a higher value promotes a positive or negative result for the class label. In this example, only the dimension of *high cost* makes it likelier for a customer to churn; the other three dimensions make it less likely for a customer to do so. ~~For convenience, we denote the tendency of a dimension with an optional arrow as a superscript (e.g. $d_4^{\uparrow}$).~~

Table 1. The features/dimensions used in the Churn example.

Dimension	Name	Description
$d_1^{\downarrow}$	Gift	Whether the customer has received a gift from the provider
$d_2^{\downarrow}$	Present	Whether the customer was present during the last organised event
$d_3^{\downarrow}$	Website	The number of times the customer logged into their a profile
$d_4^{\uparrow}$	High cost	Whether the customer is in a high-cost category

We start with an informal example that illustrates the sort of justifications produced by AF-CBA, before defining the argumentation framework formally. We take the two customers presented in Table 2 as our example CB.

Table 2. Fictional example based on the Churn data with a CB consisting of only two cases.

Customer	$d_1^{\downarrow}$	$d_2^{\downarrow}$	$d_3^{\downarrow}$	$d_4^{\uparrow}$	Label (churn)
Alice	1	0	5	0	0
Bob	1	1	3	1	1
Charlie	0	0	6	0	0

Let us now presume there is a new customer with the following values for these features:

We want a justification for the predicted outcome for this focus case. Say we have a ML model which predicts ~~that~~ Charlie will stay. Figure 1 presents this justification as a dialogue game tree, which depicts an argument between a proponent of this outcome

Table 3. Churn example focus case.

Customer	$d_1^{\downarrow}$	$d_2^{\downarrow}$	$d_3^{\downarrow}$	$d_4^{\uparrow}$	Label (churn)
Charlie	0	1	3	0	?

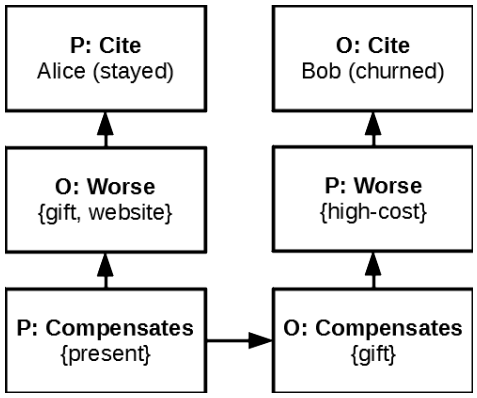

Figure 1. A fictional example of a justification generated by AF-CBA.

and an opponent arguing for the opposite. Since neither of the cases in the CB is exactly the same as the focus case, the decision is not forced and the proponent and opponent argue about the outcome using Worse and Compensates moves. The Compensates move relies on the fact that one dimension can be more important than the other. For instance, receiving a gift from your telecom provider may not be as reliable an indicator that someone will stay with that provider than for instance them attending an event organised by that provider. We assume that these relations between dimensions are known and we hereafter refer to them as the set dc. dc may be determined in any of a number of ways, such as studying feature importances in relation to the ML approach, knowledge engineering or during an explanation dialogue with a user.

The dialogue tree of Figure 1 can be read as follows:

1. Proponent: Alice stayed and her case is similar to Charlie's.
2. Opponent: Charlie's scores for $d_1^{\downarrow}$ and $d_3^{\downarrow}$ make him worse for staying than Alice.
3. Proponent: Charlie's score for $d_2^{\downarrow}$ compensates for that.
4. Opponent: Bob churned and his case is similar to Charlie's.
5. Proponent: Charlie's score for $d_4^{\uparrow}$ makes him worse for churning than Bob.
6. Opponent: Charlie's score for $d_1^{\downarrow}$ compensates for that.
7. Proponent: Charlie's score for $d_2^{\downarrow}$ compensates for *that*.

After this, the opponent has run out of possible moves to make and the proponent wins. The similarity to Alice's case has held up and acts as a justification for the prediction that Charlie will stay as well.

2.1. Argumentation framework

We will now formalise the approach depicted in the example above. ~~Before we can introduce the argumentation framework itself, we must address some more fundamental definitions and notations.~~ Let o and o' be the two possible outcomes of a case in the case

base. The variables s and $\bar{s}$ denote the two sides, meaning that $s = o$ if $\bar{s} = o'$ and vice versa. A dimension is defined as a tuple $d = (V, \leq_o, \leq_{o'})$, with value set V and two partial orderings on V, $\leq_o$ and $\leq_{o'}$, such that $v \leq_o v'$ iff $v' \leq_{o'} v$ for $v, v' \in V$. A value assignment is a pair (d, v). We denote the value x of dimension d as $v(d) = x$. Value assignments to all dimensions $d \in D$ (where D is nonempty) constitute a fact situation. A case is defined as $c = (F, outcome(c))$ for such a fact situation and an $outcome(c) \in \{o, o'\}$. In this context, a case base CB specifically refers to the set of cases with value assignments for D. We denote the fact situation of a case c as $F(c)$. For the sake of brevity in the rest of this paper, we implicitly assume that any two fact situations assign values to the same set D.

We model Horty's [5] *a fortiori* reasoning using Definitions 1 & 2, meaning that the outcome of a focus case is forced if there is a precedent with the same outcome where all their differences make the focus case even stronger for that outcome.

Definition 1 (Preference relation for fact situations)**.** *Given two fact situations F and F', $F \leq_s F'$ iff $v \leq_{v'}$ for all $(d, v) \in F$ and $(d, v') \in F'$.*

Definition 2 (Precedential constraint)**.** *Given case base CB and fact situation F, deciding F for s is forced iff CB contains a case $c = (F', s)$ such that $F' \leq_s F$.*

A fact situation could conceivably be forced for both s and $\bar{s}$, which brings us to the following definition of CB consistency:

Definition 3 (Case base consistency)**.** *A case base CB is consistent iff it does not contain two cases $c = (F, s)$ and $c' = (F', \bar{s})$ such that $F \leq_s F'$. Otherwise it is inconsistent.*

Table 4 shows another small CB based on the Churn data set. All dimensions have exactly the same values. This means that we would expect all three cases to receive the same outcome. And yet, c_3 has the opposite outcome. If c_1 were our focus case, it would be forced by both c_2 and c_3 for both outcomes, as there are no differences which make the focus case worse than those precedents. This makes the addition of c_3 to the CB a source of inconsistency.

Table 4. Example of an inconsistent CB.

Customer	$d_1^{\downarrow}$	$d_2^{\downarrow}$	$d_3^{\downarrow}$	$d_4^{\uparrow}$	*outcome*
c_1	1	1	2	0	s
c_2	1	1	2	0	s
c_3	1	1	2	0	$\bar{s}$

An explanation takes the form of an argument game for grounded semantics [11] played between a proponent and opponent of an outcome, in which they take turns to attack the other's last argument. An argument is justified if the proponent has a winning strategy, meaning the opponent runs out of moves in whatever way the opponent plays. The proponent starts the game by citing a best precedent from the case base. This is a case which has the outcome for which the proponent is arguing and has a minimal subset of relevant differences with the focus case.

The opponent can reply by playing a distinguishing move or by citing a counterexample. The proponent can reply to the latter with similar distinguishing moves, intended to show how the cited precedent has no relevant differences with the focus case. The allowed distinguishing moves are:

- $Worse(c,x)$: the focus case is on some dimensions x worse than the precedent c for $outcome(c)$.
- $Compensates(c,x,y)$: the dimensions x on which the focus case is not at least as good as the precedent c for $outcome(c)$ are compensated by dimensions y on which the focus case is better for $outcome(c)$ than c according to the compensation defintions provided by dc.

Additionally, Prakken & Ratsma describe the need for a $Transformed$ move, which is when the $Compensates$ move turns one case into another, and the need to allow the $Compensates$ move to be empty in order to state that the differences with the focus case do not matter (see the original paper by Prakken & Ratsma for a complete explanation [4]).

Determining the relevant differences between two cases is defined according to Definition 4. Definition 5 provides the criteria for a best precedent to cite.

Definition 4 (Differences between cases)**.** *Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. The set $D(c,f)$ of differences between c and f are:*

1. $D(c,f) = \{(d,v) \in F(c) \mid v(d,c) \not\leq_s v(d,f)\}$ *if* $outcome(c) = outcome(f) = s$
2. $D(c,f) = \{(d,v) \in F(c) \mid v(d,c) \in F(c) v(d,f) \mid v(d,c) \not\geq_{\bar{s}} v(d,f)\}$ *if* $outcome(c) \neq outcome(f)$

Definition 5 (Best precedent)**.** *Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases, where $c \in CB$ and $f \notin CB$. c is a best precedent for f iff:*

- $outcome(c) = outcome(f)$ *and*
- *there is no* $c' \in CB$ *such that* $outcome(c') = outcome(c)$ *and* $D(c',f) \subset D(c,f)$.

For example, in Table 4, $outcome(c_1) = outcome(c_2)$ and $outcome(c_2) \neq outcome(c_3)$. $D(c_2,c_1) = \emptyset$ since there are no differences between c_1 and c_2. If it were true that $D(c_3,c_1) \neq \emptyset$, than c_3 would be a case with the opposite outcome and relevant differences with the focus case, disqualifying c_2 as a best precedent. However, since Table 4 shows that $D(c_3,c_1) = \emptyset$, c_2 is indeed the best precedent for c_1.

This brings us to the following definition for the AF-CBA framework:

Definition 6 (Case-based argumentation framework)**.** *Given a finite case base CB, a focus case $f \notin CB$, and definitions of compensation dc, an abstract argumentation framework AAF is a pair $< A, attack >$, where:*

- $A = CB \cup M$,
  with $M = \{Worse(c,x) \mid x \neq 0$ *and* $x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}\} \cup$
  $\{Compensates(c,y,x) \mid y \subseteq \{(d,v) \in (f) \mid v(d,c) <_{outcome(f)} v(d,f)\},$
  $x = \{(d,v) \in F(f) \mid v(d,f) <_{outcome(f)} v(d,c)\}$ *and y compensates x according to* $def\} \cup$
  $\{Transformed(c,c') \mid c \in CB$ *and c can be transformed into* c' *and* $D(c',f) = 0\}$
- *A attacks B iff:*
  - $A, B \in CB$ *and* $outcome(A) \neq outcome(B)$ *and* $D(B,f) \not\subset D(A,f)$;
  - $B \in CB$ *with* $outcome(B) = outcome(f)$ *and A is of the form* $Worse(B,x)$;

* *B is of the form* $Worse(c,x)$ *and A is of the form* $Compensates(c,y,x)$*;*
* $B \in CB$ *and* $outcome(B) \neq outcome(f)$ *and A is of the form* $Transformed(c,c')$*.*

3. CB inconsistency

As we argued in Section 1, CB consistency is not always a safe assumption to make. In order to meaningfully apply the explanation method to an inconsistent CB, we must mitigate the problem that a focus case can be forced for both outcomes. Instead, the inconsistency that causes this problem must prevent the focus case from being forced. That is, forcing a case should only happen when the CB is fully consistent. When there is consistency, a precedential case has a stronger backing when cited and can thus force the outcome; if there is inconsistency, it has less backing and thus cannot. We therefore introduce the concept of 'authoritativeness', by which we mean that, given any case $c \in CB$ with outcome s, the authoritativeness $\alpha(c)$ numerically expresses (normalised between 0 and 1) the degree to which the rest of the CB supports the citing of c for s. The intuition behind authoritativeness is that whereas the a fortiori rule applied to a consistent CB can be expressed as the phrase 'cases like this always receive outcome o,' our idea of authoritativeness changes this phrase to 'cases like this *usually* receive outcome o'—where 'usually' has to be quantified in some manner. Since $\alpha(c)$ is a number, we can have a total ordering $\leq$ on the authoritativeness of cases. In this section, we formulate possible expressions for authoritativeness.

Table 5 is another instance of our Churn example. This time, c_1 and c_2 should receive a higher value for $\alpha(c)$ than c_3 and c_4, since c_1 and c_2 have the same outcome, whereas c_3 and c_4 are inconsistent with each other.

Table 5. Example of a CB with two cases that are consistent with each other and two cases which contradict each other.

Customer	$d_1^{\uparrow}$	$d_2^{\uparrow}$	$d_3^{\uparrow}$	$d_4^{\downarrow}$	*outcome*
c_1	1	1	0	0	s
c_2	1	1	0	0	s
c_3	1	1	5	0	s
c_4	1	1	5	0	$\bar{s}$

First of all, the definition of best precedent has to be modified to reflect the additional criterion of maximising the authoritativeness:

Definition 7. *(Best authoritative precedent) Let CB be a case base and let* $c = (F(c), outcome(c))$ *and* $f = (F(f), outcome(f))$ *be two cases, where* $c \in CB$ *and* $f \notin CB$*. c is a best precedent for f iff:*

- $outcome(c) = outcome(f)$*,*
- *there is no* $c' \in CB$ *such that* $outcome(c') = outcome(c)$ *while* $D(c',f) \subset D(c,f)$ *and* $\alpha(c', outcome(c)) > \alpha(c, outcome(c))$*.*

In order to quantify authoritativeness, we require expressions of agreement and disagreement between a precedent and the rest of the CB:

Definition 8. *(Agreement) Let CB be a case base and let* $c = (F(c), outcome(c))$*, where* $c \in CB$*. The agreement* $n_a(c)$ *is defined as:*
$n_a(c) = \mid \{c' \in CB \mid outcome(c') = outcome(c) \text{ and } D(c,c') = \emptyset\} \mid$

Definition 9. *(Disagreement) Let CB be a case base and let* $c = (F(c), outcome(c))$*, where* $c \in CB$*. The disagreement* $n_d(c)$ *is defined as:*
$n_d(c) = \mid \{c' \in CB \mid outcome(c') \neq outcome(c) \text{ and } D(c,c') = \emptyset\} \mid$

We understand $n_a(c)$ as the number of cases which have the same outcome as the precedent case without any relevant differences. Similarly, $n_d(c)$ is the number of cases which have the opposite outcome yet also have no relevant differences. The agreement $n_a(c)$ has at least a value of 1 due to c itself being a member of the CB. Since c cannot also be a member of the CB with the opposite outcome, the disagreement can have a value of 0.

Using these numbers, we can formulate expressions of authoritativeness. Exactly how the level of agreement relates to authoritativeness is not self-evident, as various expressions may have equal merit. For example, given a case c in case base CB, we could express the authoritativeness $\alpha(c)$ as the percentage of cases with the same differences with the focus case with the same outcome, relative to those with the opposite (1). The intuition behind this is that the ratio of agreement and disagreement expresses the strength of citing a precedent c.

Looking back at Table 4, c_1 is backed by one other case, while c_3 does not support c_1 due to its opposite outcome. So in that situation, $\alpha(c_1, s) = 2/3$.

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \tag{1}$$

However, this overlooks any intuitive understanding of authoritativeness which stems from the absolute number of cases that can act as precedents (2). The intuition behind this is that obscure cases are less authoritative than common ones, which naturally form a larger group of precedents. In Table 6, c_1 is backed by two cases (c_2 and c_3), c_4 is backed by c_5 and c_7 is backed by none. We divide by $\mid CB \mid$ to normalise the expression between 0 and 1, which was not necessary with the previous expression. So for example $\alpha(c_1, s) = 3/7 \approx 0.429$.

$$\alpha(c) = \frac{n_a(c)}{\mid CB \mid} \tag{2}$$

Both (1) and (2) would appear to have some merit, at least considering their intuitive understanding. As such, using a combination of the two seems even more intuitive. There are multiple ways of achieving this. One possibility is to take the product of these two alternatives, essentially using the relative number of similar cases in agreement with the focus case as a multiplicative weight factor for the absolute number of cases (3). When there is no inconsistency, this is equal to (2).

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \cdot \frac{n_a(c)}{\mid CB \mid} \tag{3}$$

Table 6. Example of an inconsistent CB showcasing three different levels of support.

Customer	$d_1^{\uparrow}$	$d_2^{\uparrow}$	$d_3^{\uparrow}$	$d_4^{\downarrow}$	*outcome*
c_1	1	1	0	0	s
c_2	1	1	0	0	s
c_3	1	1	0	0	s
c_4	1	1	2	0	s
c_5	1	1	2	0	s
c_6	1	1	2	0	$\bar{s}$
c_7	1	1	15	0	s

Instead of multiplying the two in this manner, (1) and (2) can be combined as a harmonic mean of the two (4). This introduces a parameter β, which represents the relative importance of one expression over the other in the computation of the harmonic mean. The added advantage of this is that (1) could be considered twice as important than (2), for instance. At a value of $\beta = 1$, the two are treated as equally important. In that case, for example, c_1 from Table 6 would receive $\alpha(c_1, s) = \frac{3/3 \cdot 3/7}{3/3+3/7} = 0.3$.

$$\alpha(c) = (1+\beta^2) \cdot \frac{\frac{n_a(c)}{n_a(c)+n_d(c)} \cdot \frac{n_a(c)}{|CB|}}{\frac{n_a(c)}{n_a(c)+n_d(c)} + \frac{n_a(c)}{|CB|}} \tag{4}$$

How desirable each expression is, is difficult to say. In the next section, we attempt to answer this question through experimentation.

References

[1] Lipton Z. The mythos of model interpretability. Communications of the ACM. 2016;61:96-100.

[2] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Computing Surveys. 2018;51(5):93:1-93:42.

[3] Miller T. Explanation in artificial intelligence: insights from the social sciences. Artificial Intelligence. 2019;267:1-38.

[4] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: formalisation and experiments. Argument & Computation. 2021;Preprint(Preprint):1-36. Publisher: IOS Press.

[5] Horty J. Rules and reasons in the theory of precedent. Legal Theory. 2011;17:1-34.

[6] Horty J. Reasoning with dimensions and magnitudes. Artificial Intelligence and Law. 2019;27(3):309-45.

[7] Aleven V. Teaching case-based argumentation through a model and examples; 1997.

[8] Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv:210314749 [cs, stat]. 2021. ArXiv: 2103.14749.

[9] Frenay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems. 2014;25(5):845-69.

[10] Telco Customer Churn;. Available from: `https://www.kaggle.com/blastchar/telco-customer-churn`.

[11] Modgil S, Caminada M. Proof Theories and Algorithms for Abstract Argumentation Frameworks. In: Simari G, Rahwan I, editors. Argumentation in Artificial Intelligence. Boston, MA: Springer US; 2009. p. 105-29.