Practical Business Analytics – Jupiter Group Project

Bechelet, Amit ab02527; Chenthil Arumugam, Raj Kannan rc00981; Hay, Graham gh00497; James, Donald dj00315; Parol, Hisham hp00540; Smith, Michael ms02394

Analysis of Investment Dispute Outcome for International Investors

Table of Contents

1 Investment Treaty Cases - Dataset Description

https://investmentpolicy.unctad.org/investment-dispute-settlement?id=12&name=austria&role=investor

This is a UN related website which provides data on all pending and concluded Investor v State investment treaty arbitration cases. These are cases brought under international treaties between states whereby an investor from one state (the Home State) can sue another State (the Host State) in which it has made an investment and then been subject to unlawful treatment sometimes amounting to expropriation. For example, a case might involve a Canadian mining company that invested in Mongolia having obtained a uranium mining licence which was then unfairly revoked and say awarded to a Russian mining company.

As at 2 November 2019, 981 such cases were listed on this website. Of these there were 640 listed as concluded cases so the sample size (even allowing for missing data items and outliers) should be sufficient for training and testing. Once we excluded Pending Cases and those for which outcome data was not available, then the dataset was 540 concluded cases.

This dataset was not available in a readily downloadable format when the project commenced. However, the group used various techniques to transform the data into a tractable CSV file described further below. (Note, on 27 November 2019 UNCTAD announced the release of all the information on the website in Excel format. The group did not have time to work with this new more formalised dataset).

In addition, the group complemented the above data with a second UN dataset which is downloadable as a spreadsheet:
http://hdr.undp.org/en/composite/HDI

This dataset provides a measure of human development for each sovereign State known as the Human Development Index (HDI). HDI is a summary measure of achievements in three key dimensions of human development:

- a long and healthy life;
- access to knowledge; and

- a decent standard of living.

The HDI is the geometric mean of normalised indices for each of the three dimensions and returns a value between 0 and 1 for each State, in fact between approx. 0.35 lowest to 0.95 highest. This data was added to the master CSV file.

2 Problem Definition

Is it possible to build models/identify key exogenous features from the chosen datasets (UNCTAD Model) which have some predictive power as to case outcomes?

The ***core hypothesis*** that the group sought to test is the following: individual factual case analysis should be fundamental and determinative when deciding on the likely outcome of an investment treaty case.

The UNCTAD model using just the identified datasets explicitly recognises that no direct idiosyncratic factual enquiry is made into the relative strengths of each underlying case, relying only on the identified features. Therefore, the model enables an assessment of the variance in case outcomes attributable to the UNCTAD Model features versus the sample mean, and thus also an assessment of the unexplained variance in case outcomes attributable to idiosyncratic case features. In other words, the model should provide guidance on how important a review of the simple exogenous factors should be in making case assessments. Moreover, at the very least the model should provide guidance as to how resources should be committed in determining which cases *a priori* appear stronger and worth selecting for deeper manual review.

The central task for the group therefore was to build and compare models that attempt to predict the win/lose and/or damages outcome of cases based on some, or all, of the following exogenous features (UNCTAD Model Features):

1. State of Claimant/HDI score;
2. Respondent State/HDI score;
3. Identity of Judges (arbitrators) – specifically, do the cases involve the most frequently appointed arbitrators by Claimant, Respondent and as President?
4. Is the claim for Direct Expropriation of the investment, or only the lower standards e.g. Unfair and Inequitable Treatment; Indirect Expropriation;
5. Which Rules were applicable to the arbitration – ICSID (World Bank related), UNCITRAL (UN Related) or those of some other institution such as the International Chamber of Commerce;

6. The amount of damages claimed;
7. From which Industry Sector does the claim originate – specifically is it from Mining; Electricity and Gas; Construction, or some other less heavy industry?

In addition to the core hypothesis the following ***additional hypotheses*** were selected as potentially suitable for consideration depending on which modelling approaches would be adopted:

1. Respondent States with a lower HDI ranking are penalised by arbitrators disproportionately when compared to higher ranking States;
2. Claimants from higher ranking HDI States receive preferential treatment;
3. Claims in heavy industry sectors (Mining, Power, Construction) outperform other cases;
4. Direct expropriation cases perform better than other cases brought under the lower treaty breach standards;
5. Amount of damages claimed is positively correlated to amount of damages awarded;
6. Choice of (some of) the "usual suspect" Arbitrators by Claimant, Respondent or as President (versus choosing another) can have an impact on case outcome.

3 Data Preparation, Pre-processing, Integration and Exploration

3.1 Data Extraction

As noted above, at the time of data extraction, the UNCTAD web site did not provide a mechanism for download of the case data, so this had to be achieved by a web scraping process. The first step was manually to view the Damages tab of the Investment Dispute Settlement Navigator page (https://investmentpolicy.unctad.org/investment-dispute-settlement) and then select the 'All' check box for the three data filters: Sought by Claimants, Awarded by Tribunal and Agreed in Settlement.

The navigator page uses ajax (Asynchronous Java script and Xml) to request data as the user scrolls down the data table. So, it was necessary manually to scroll down the table until it was populated with a data set of 714 cases (which included some cases that were still pending). This was done using the Chrome web browser which has a set of development tools that enable the full html DOM (Document Object Model) to be viewed as html text. This was copied and saved as a UTF-8 encoded html text file (Damages.html). The html file can be loaded and parsed using the 'rvest' R library and written to a raw UTF-8 encoded csv file (Damages.csv) ready for pre-processing.

See R function JextractDamages in the source code.

3.2 Data Cleaning

This dataset needed a substantial amount of work to clean a number of the fields in order to make them coherent and normalised for analysis purposes as well as interpretation. These fields consisted of free text with little formatting to aid transformation. The task to transform these fields used several custom functions, which deal with the complexity of the text. Ultimately the functions were written generically in order to be utilised in multiple cases, for example 'rev_char' & 'add_separator'. The data cleaning task consumed more resources than anticipated, which impacted the time remaining for subsequent tasks. See R function 'JFunctions.R' in the source code for all data cleaning functions used. In particular, the following fields were cleaned:

3.2.1 Amount Fields

The raw amount fields (‘Amount claimed’ & ‘Amount awarded’) consisted of string values, sometimes in multiple currencies. USD was always included, so it was decided to remove the amount given in non-USD currencies and retain only the USD values. In order to do this, it was necessary to first detect the section of the field with the USD values. Following that the string was split, and the additional currency removed. Once this was done the string was converted to numeric value.

3.2.2 Arbitrators

This was a single field with multiple string values, which included the full history of previous values. Ultimately 3 separate values needed to be extracted from this field: Arbitrator on behalf of the Claimant (column ‘aClaimantC’); Arbitrator on behalf of the Respondent (column ‘aRespondentC’); Arbitrator presiding as President (column ‘aPresidentC’). The previous history (‘replaced’ value) was removed, and the three values extracted to their relevant columns. Furthermore, there were some non-English characters used in some of the names. In order to preserve these characters uncorrupted it was necessary to encode the file correctly.

3.2.3 Home State Field – Claimant State

Some instances of this field had multiple values, which impeded classification. It was therefore decided, with some domain knowledge, that it was reasonable that the first value should be taken. Generally, the first named state will be the parent state where the capital invested originates, or where the investor business is headquartered and the second will be the state of some intermediate holding company. Occasionally the claimants will be a group, from similar very high HDI states and the first named will form a suitable proxy for the average HDI of the group.

3.2.4 Update HDI’s

The respondent state and home state values were not aligned with the state values in HDI. This function compared the value and updated with that from HDI. At the same time the corresponding HDI value was added to the main table.

3.3 UNCTAD Model Features - Final Variable Selection

The variables described below were selected for the models and are retained in a Data frame for the modelling process. See R function JDataPreparation in the source code.

Categorical columns have been encoded using binary (0 or 1) or 1-hot encoding. In 1-hot encoding, a column is created for the categorical values and the value is indicated by a 1 or 0. All columns having a value of zero indicates a value of 'Other'.

Arbitral Rules - Categorical 1-hot-encoded

Value	Data Frame Column – 1 hot encoded	Count
ICSID	Arbitralrules_ICSID	274
UNCITRAL	Arbitralrules_UNCITRAL	182
ICC	Arbitralrules_ICC	9
SCC	Arbitralrules_SCC	32
ICSIDAF	Arbitralrules_ICSIDAF	38

Breaches Alleged – Categorical Binary encoded

Value	Data Frame Column – binary encoded
Contains 'Direct expropriation'	breaches

Economic Sector

Value	Data Frame Column – 1 hot encoded	Count
Primary: B - Mining	EconmicSector_PrimaryBMining	81
Tertiary: D – Electricity	EconmicSector_TertiaryDElectricity	101
Tertiary: F – Construction	EconmicSector_TertiaryFConstruction	46

Arbitrators – President

Value	Data Frame Column – 1 hot encoded	Count
Kaufmann-Kohler, G.	aPresidentC_KaufmannKohlerG	25
Armesto, J.	aPresidentC_ArmestoJ	14
Veeder, V. V.	aPresidentC_VeederVV	15
Fortier, L. Y.	aPresidentC_FortierLY	21
Tercier, P.	aPresidentC_TercierP	12

Arbitrators – Claimant

Value	Data Frame Column – 1 hot encoded	Count
Brower, C. N.	aClaimantC_BrowerCN	34
Alexandrov, S. A.	aClaimantC_AlexandrovSA	15
Fortier, L. Y.	aClaimantC_FortierLY	16
Orrego Vic	aClaimantC_OrregoVic	15
Grigera	aClaimantC_Grigera	8
Beechey, J	aClaimantC_BeecheyJ	11
Hanotiau, B.	aClaimantC_HanotiauB	10

Arbitrators – Respondent

Value	Data Frame Column – 1 hot encoded	Count
Stern, B.	aRespondentC_SternB	52
Thomas, J. C	aRespondentC_ThomasJC	23
Sands, P	aRespondentC_SandsP	16
Douglas, Z.	aRespondentC_DouglasZ	9
Landau, T.	aRespondentC_LandauT	14

Claimant HDI

This is the HDI Index of the country of the investor. It is a value between 0 and 1 and so does not require normalisation. Any NULL or NA values are replaced by the mean value but in fact all values are present.

Respondent HDI

This is the HDI Index of the Respondent State. It is a value between 0 and 1 and so does not require normalisation. Any NULL or NA values are replaced by the mean value but in fact all values are present.

Success

This is the class label derived from Outcomeoforiginalproceedings to be predicted:

1 - where ‘Decided in favour of investor’ or ‘Settled’

0 - otherwise

4 Section 4 - Selection of Modelling approaches

Having obtained and pre-processed the dataset, the group considered the most suitable modelling approaches to analyse the various hypotheses. It was observed that the dataset contained a mix of continuous numerical and discrete categorical data, so would require handling with regression and classification techniques, namely:

- Classification of overall outcome (Successful Claim or Not, i.e. Win/Lose);
- Regression analysis on damages awarded/claimed given successful outcome.

To implement the analysis the following modelling techniques were considered:

- Regression – Linear; Non-linear;
- Classifications – Decision Tree; Boosted DT; Random Forest; Deep Neural Network; Multi-layer perceptron.

4.1 Simple Regression on the continuous data

On the continuous data, the group set out to explore whether certain expected simple correlations (most likely linear) existed, namely, in the 181 cases (out of 540 concluded cases) where ***damages were awarded***. Specifically, the group postulated the following questions (subject to time constraints) for consideration:

- Is there correlation between Claimant HDI and the amount of damages?
- Same analysis for Respondent HDI.
- Also analyse if there is a correlation between damages claimed and damages awarded.
- Renormalise Awarded Amount by dividing by Claimed Amount to give Damages Ratio, then rerun the above analyses for Claimant HDI, and Respondent HDI versus Damages Ratio.

These questions were approached through various simple regression analyses in R and the results are set out below in the Model Output section. The intention was to use the output to analyse whether additional hypotheses 1, 2 and 5 were supported.

4.2 Discrete categorical and continuous data

On the combined discrete and continuous data, the group set out to create models to explore whether the core and certain of the additional hypotheses (numbers 3, 4 and 6) were supported by the model output on the 540 concluded cases. (NB. the group excluded cases where data was not available or case pending - 174 of 714 Outcomes).

The first categorical model selected was to create a Decision Tree where the case outcome (from total sample of 540 cases) was to be predicted. The output variable would be either:

- ***Success*** – settled or decided in favour of Investor (280 of 540 outcomes); or
- ***Failure*** – discontinued, decided in favour of neither party or decided in favour of State (260 of 540 outcomes).

The group used a hold-out method of splitting the training and test data but subsequently enhanced this with K-fold validation. Given the data set was 540 cases, this suggested that we create 6-fold validation subsets with 90 cases in the Test Set (in each 6-fold) and 450 cases in each training set. The group noted that the sample split almost evenly between Success and Failure, (51.852% versus 48.148%). The sets were also chosen to ensure each Training Set of 450 maintained the sample balance between Success and Failure.

The decision tree was then constructed to have the following nodes (see also previous section – 3.3 Final Variable Selection):

1. Choice of Claimant Arbitrator (6 values being the 5 most commonly appointed plus other: Brower C, Alexandrov S A, Fortier L Y, Orrego Vicuna F, Grigera Naon H A, Other);
2. Choice of Respondent Arbitrator (6 values being the 5 most commonly appointed plus other: Stern B, Thomas J C, Sands P, Douglas Z, Landau T, Other);
3. Choice of President Arbitrator (6 Values being the 5 most commonly appointed plus other: Kaufmann-Kohler G, Fernandez-Armesto J, Veeder V V, Fortier L Y, Tercier P, Other);
4. HDI Score Investor State (continuous value);
5. HDI Score Respondent State (continuous value);
6. Industry Sector (4 Values being the 3 “heavy” industries plus other: Mining, Electricity etc, Construction, Other);

7. Arbitral Rules (6 Values being the 5 most commonly selected applicable rules plus other: ICSID, ICSID AF, UNCITRAL, SCC, ICC, Other);
8. Claim for the highest standard of breach being Direct Expropriation included or not (2 Values: Yes, No).

The group then ran additional analyses using enhanced Decision Tree methods (Boosted DT; Random Forest) plus Deep Neural Network and Multi-layer perceptron models. All utilised K-Fold validation.

5 Section 5 - Model Output (Results) and Assessment

5.1 Model Output on continuous only data

The group ran each of the following regression analyses:

- Analysis of correlation between Claimant HDI and the amount of damages.
- Analysis of correlation between Respondent HDI and the amount of damages.
- Analysis of correlation between damages claimed and damages awarded.
- Renormalise Amount Awarded by dividing by Amount Claimed to give Damages Ratio, then rerun the above analyses for Claimant HDI, and Respondent HDI versus Damages Ratio.

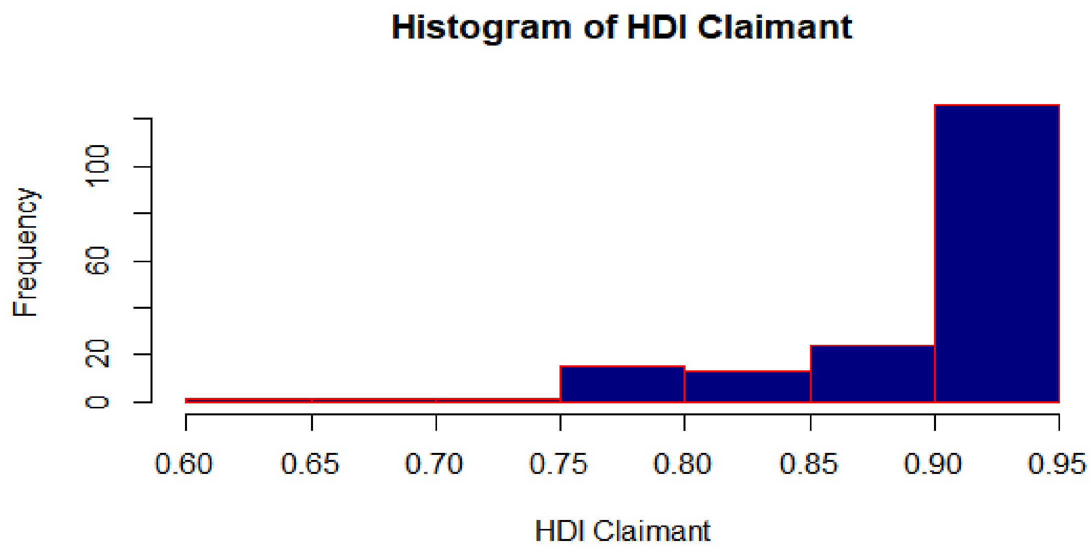
The Linear Regression Models used were very simple but enabled a quick grasp of the absence or existence of basic relationships in the continuous data, and therefore an assessment of additional hypotheses 1, 2 and 5. The regression model effectiveness in addressing the hypotheses is discussed in Section 6 – Evaluation below. In terms of technical assessment, the group simply noted the usual shortcomings with linear regression models applied here, namely that the model was very effective in identifying the existence (or not) of linear (and log-linear) relations but was unable to provide insights into the deeper hypotheses being considered, especially the core hypothesis.

The regression sample was 181 cases from which upper quartile outliers were removed using an interquartile range derived formula. In practice this meant that cases with damages awarded in excess of $2 billion (6 cases) were excluded from certain calculations.

The respective graphs output from the model for each of the above regression analyses are set out below along with some simple histograms, which illustrate the prevailing trend that Claimant HDI is greater than Respondent HDI in cases where damages are awarded, but that Respondent HDI still typically exceeds 0.65. In other words, damages were generally awarded against Respondent states with a reasonable standard of human development.

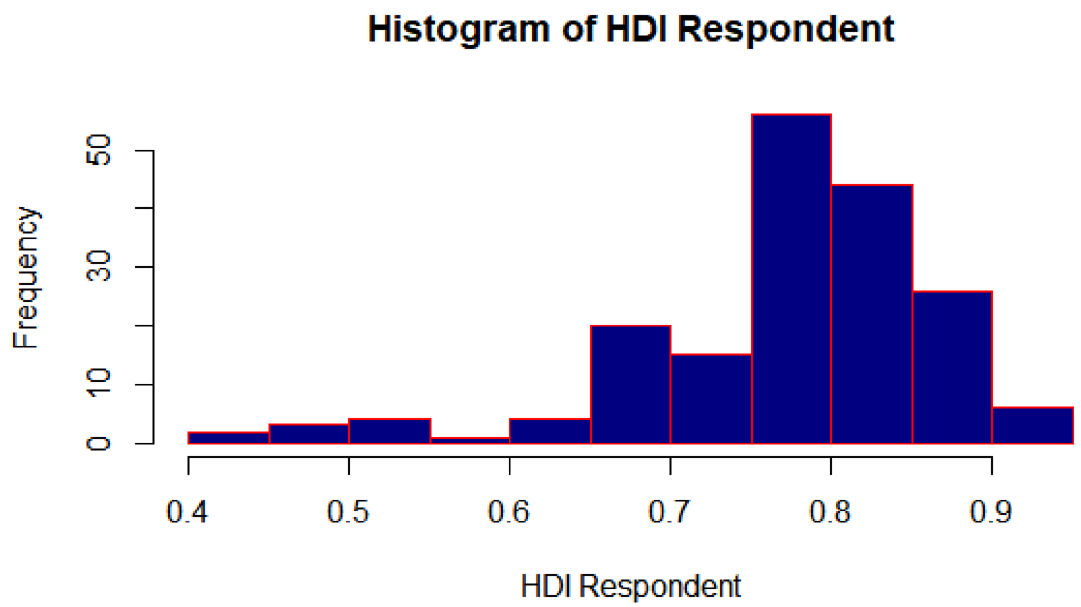
Histogram HDI Claimant

The group noted that the frequency of disputes filed by the claimant states is high with HDI value greater than 0.75



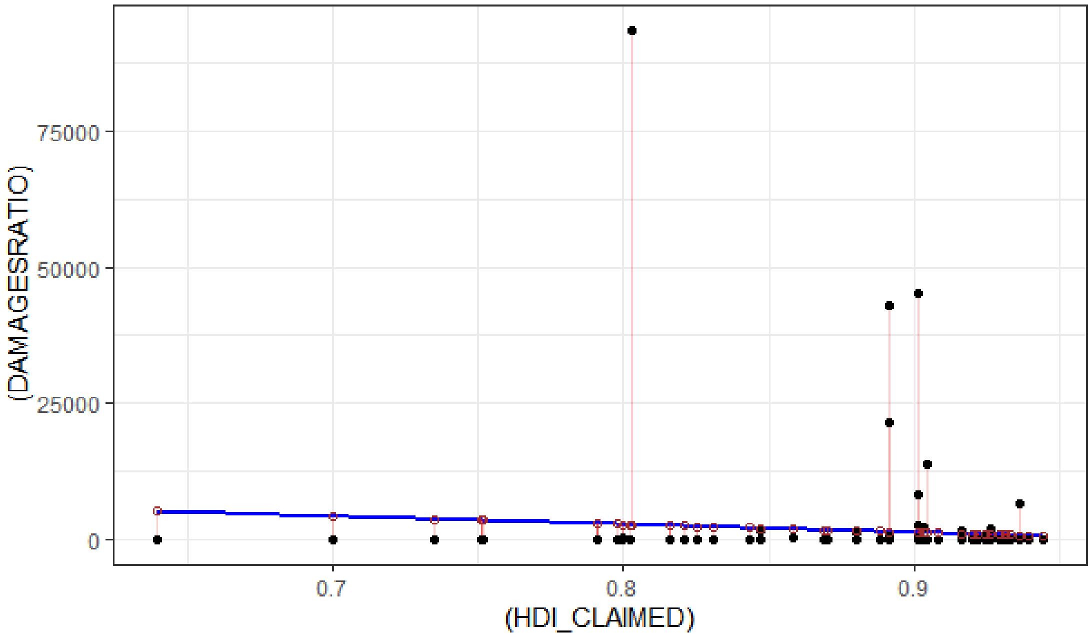
Histogram of HDI Respondent

Virtually all the sample cases are filed against states whose HDI value is greater than 0.65. The histogram below clearly depicts that the majority of disputes are filed against states with HDI values between 0.75 and 0.9.



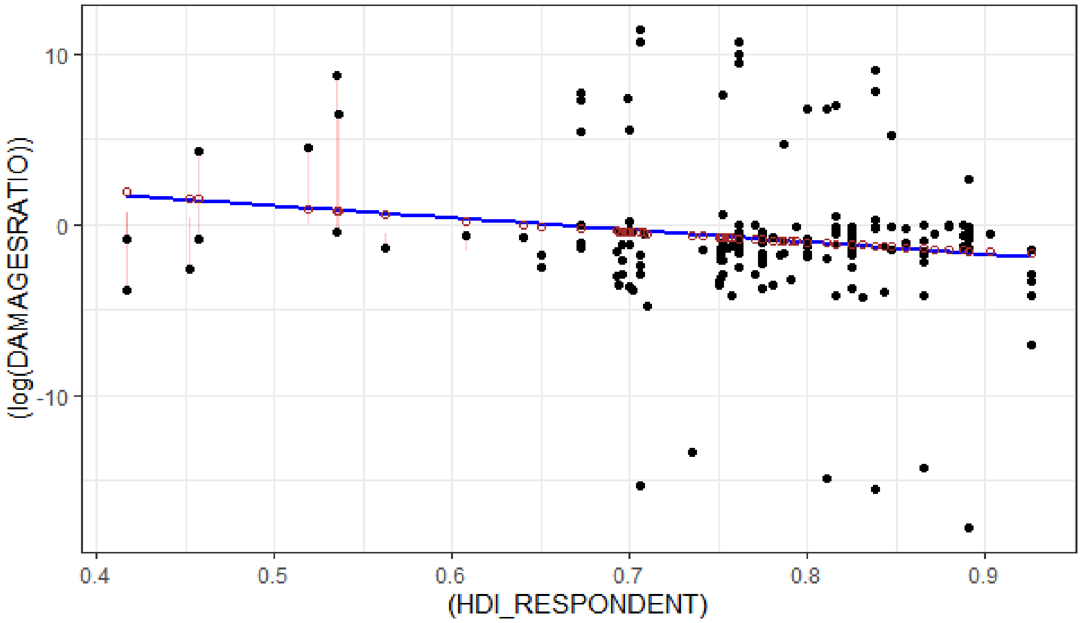
Correlation between Claimant HDI and Damage Ratio

The R^2 value obtained from the regression analysis is 0.00862, so the correlation is essentially zero when comparing Claimant HDI with the Damage Ratio. This is also self-evident from the graph below.



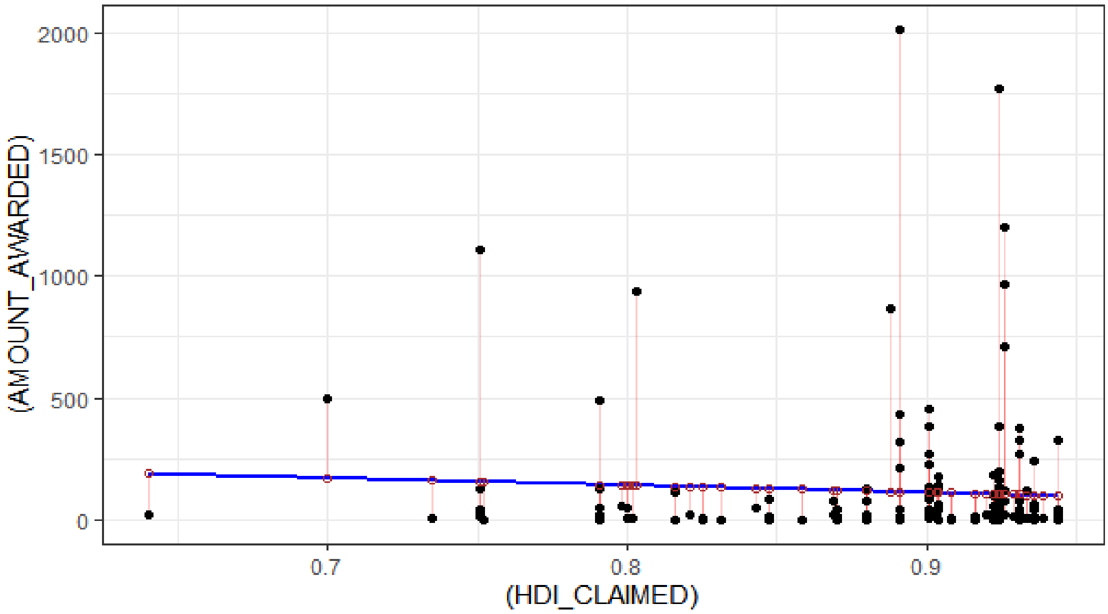
Correlation between Damage ratio and Respondent HDI

The R^2 value obtained from the regression analysis is 0.02347, so the correlation is very low when comparing Respondent HDI with the Damage Ratio. This is also self-evident from the graph below.



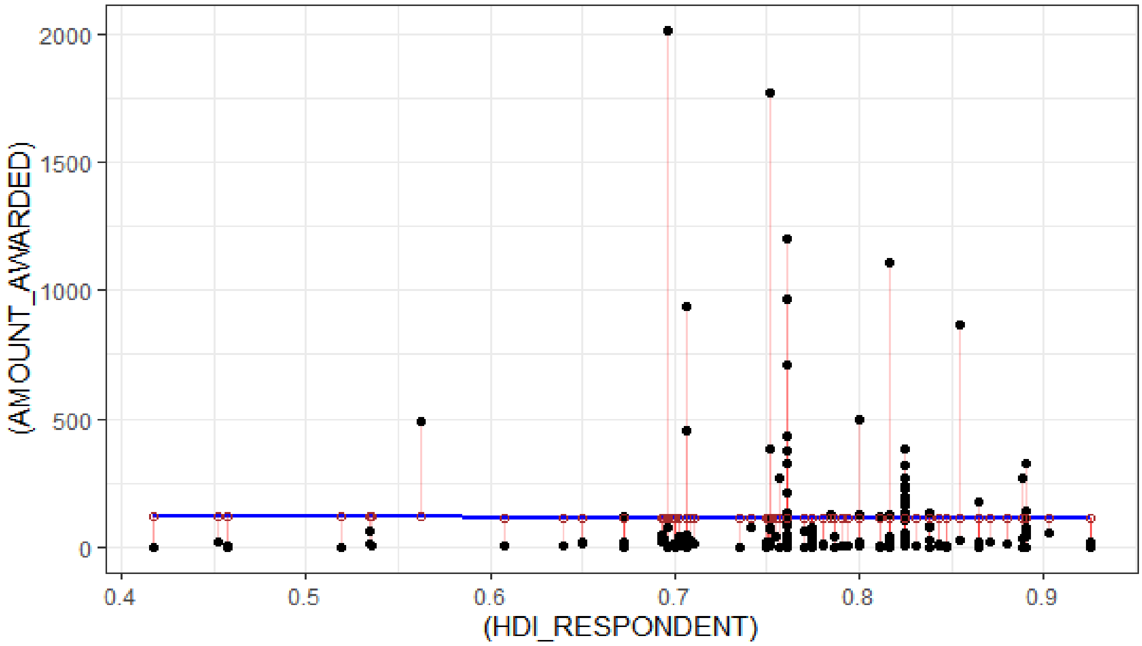
Linear Regression between Claimant HDI and Amount awarded

The R^2 value when comparing claimant HDI against Amount Awarded is 0.02347. Again, this depicts low correlation between the variables.



Linear Regression between Respondent HDI and Amount Awarded

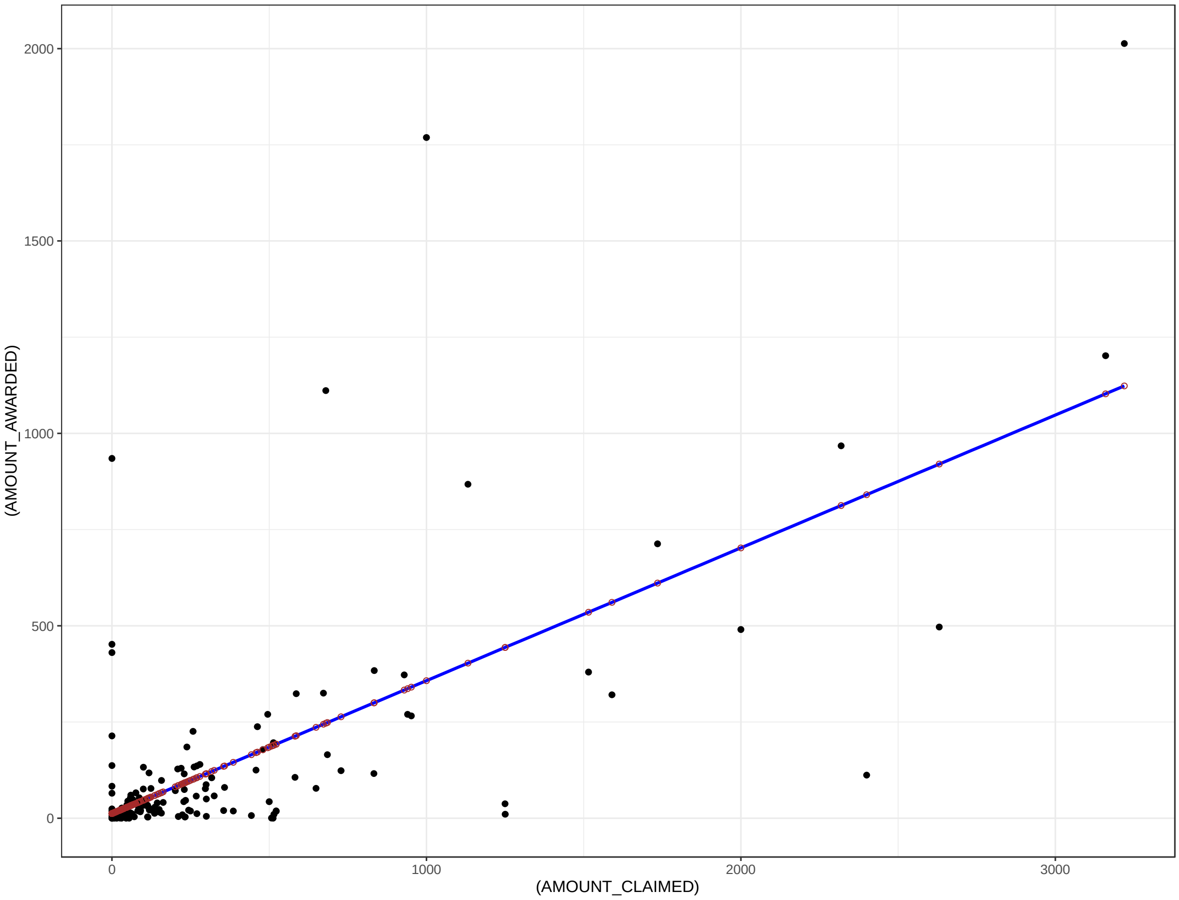
The R^2 value when comparing Respondent HDI against Amount Awarded is extremely small - $2.816e^{-05}$. This depicts essentially zero correlation between the variables.



2000
1500
1000
500
0
(AMOUNT_AWARDED)
0.4
0.5
0.6
0.7
0.8
0.9
(HDI_RESPONDENT)

Linear Regression between Amount Claimed and Amount Awarded

The R^2 value when comparing Amount Claimed against Amount Awarded is 0.4793. This depicts significant correlation between these variables. In addition, the slope of the line is 0.345.



5.2 Model Output on Discrete categorical and continuous data

As noted above, the group explored whether the core and some of the additional hypotheses (numbers 3, 4 and 6) were supported by the model output on the 540 concluded cases. For reference, these hypotheses were:

- ***Core hypothesis***: individual factual case analysis should be fundamental and determinative when deciding on the likely outcome of an investment treaty case, and/or can it be said that the UNCTAD model appears to explain (some of) the variance in case outcomes?
- Claims in heavy industry sectors (Mining, Power, Construction) outperform other cases;
- Direct expropriation cases perform better than other cases brought under the lower treaty breach standards;
- Choice of (some of) the "usual suspect" Arbitrators by Claimant, Respondent or as President (versus choosing another) can have an impact on case outcome.

To implement the analysis on the categorical / continuous data the following non-linear classifier modelling techniques were used:

1. Decision Tree;
2. DT with K-fold validation;
3. Boosted DT;
4. Random Forest;
5. Deep Neural Network;
6. Multi-layer perceptron.

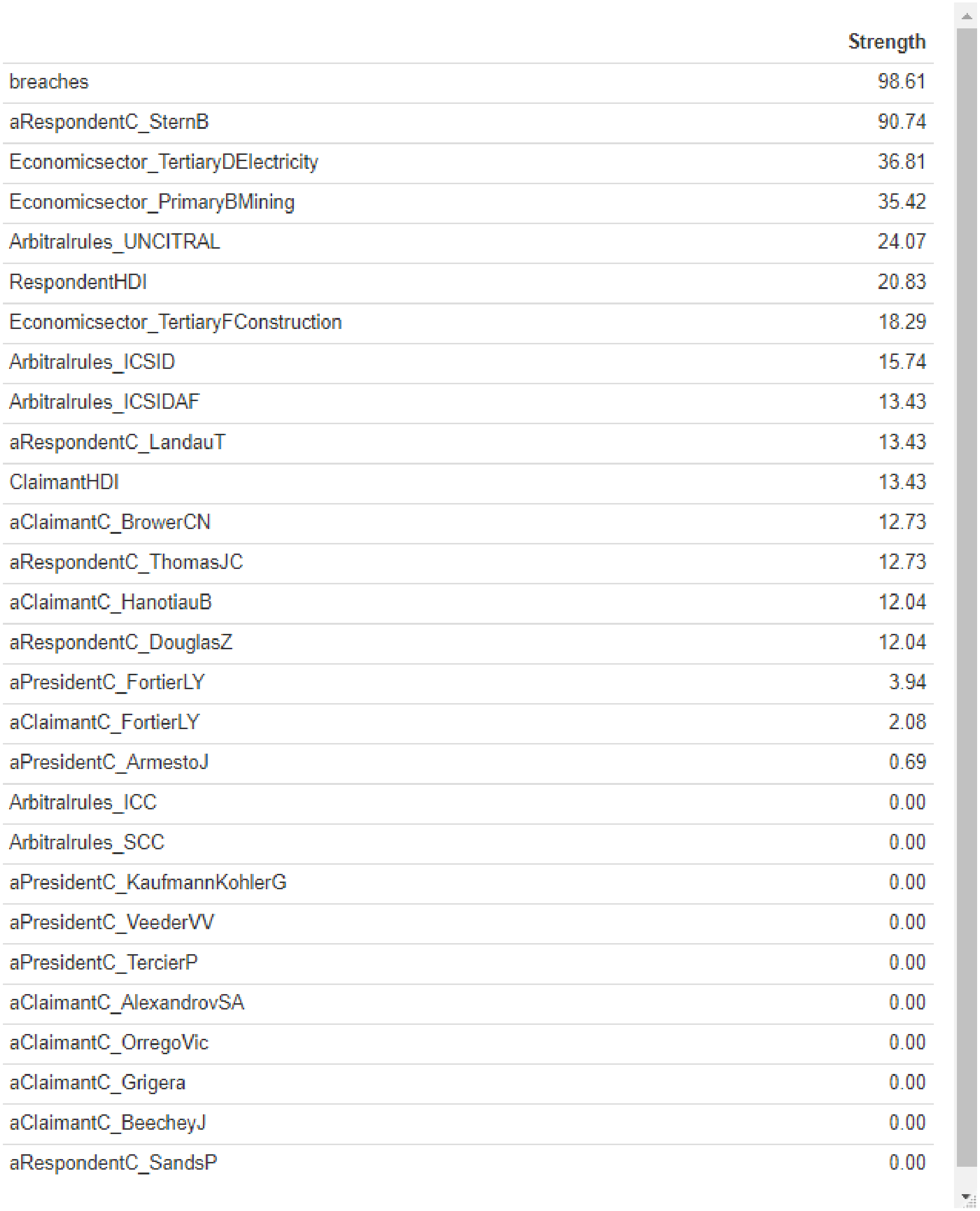
The individual results are set out below, following which we provide a summary evaluation.

5.2.1 Decision Tree – Holdout Method

A decision tree C5.0 was run using a holdout method of 80/20 training to test ratio, which were compromised of 432 cases with 29 attributes. 13 rules were derived using the full set of variables. The resulting ROC determined a threshold of 0.4854 with TPR: 63.16% and FPR: 41.18%.

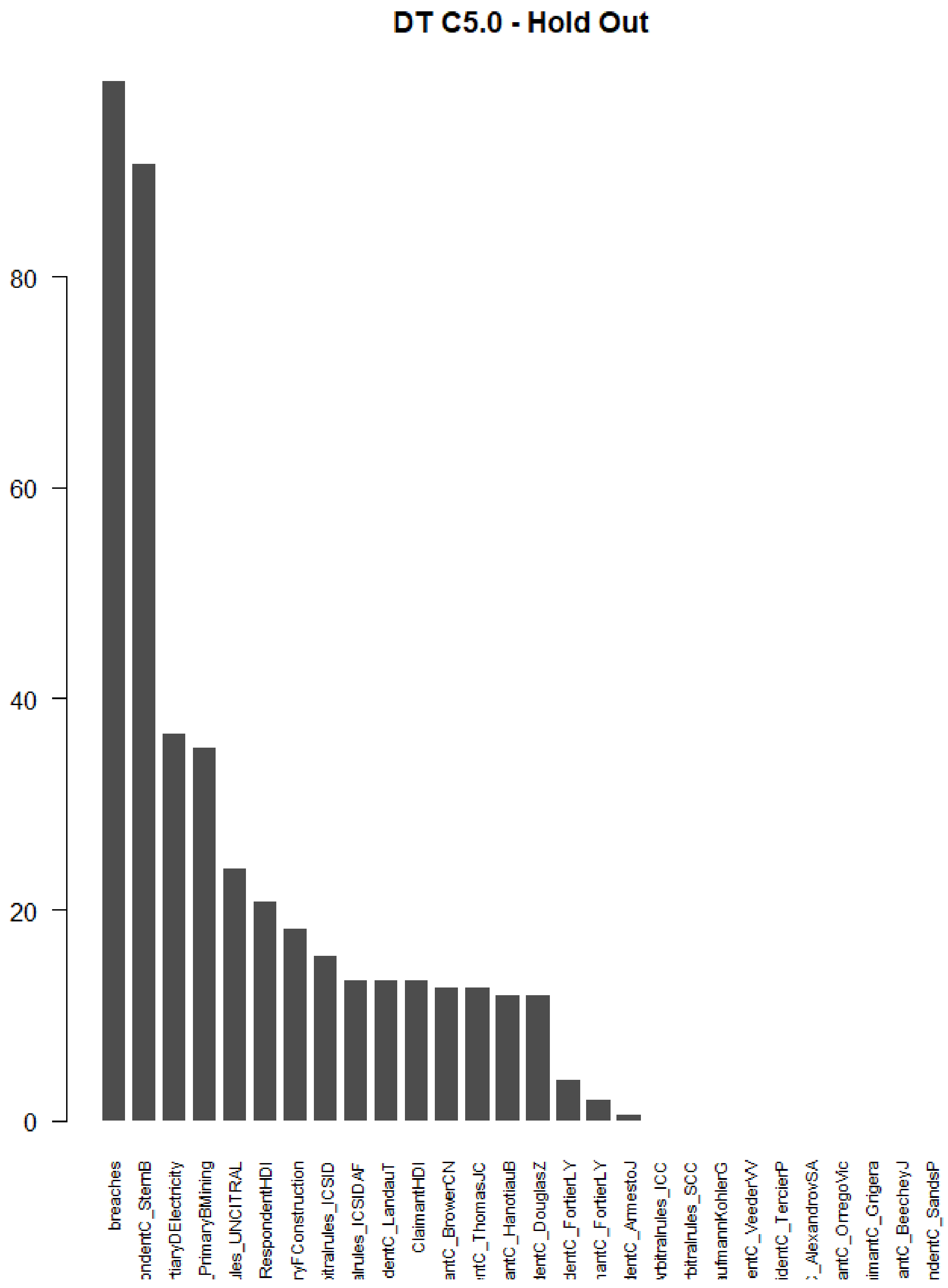
Measure	Averaged Value
TP	36
FN	21
TN	30
FP	21
Accuracy	61.11
pgood	63.16
pbad	58.82
FPR	41.18
TPR	63.16
TNR	58.82
MCC	0.22
Threshold	0.31

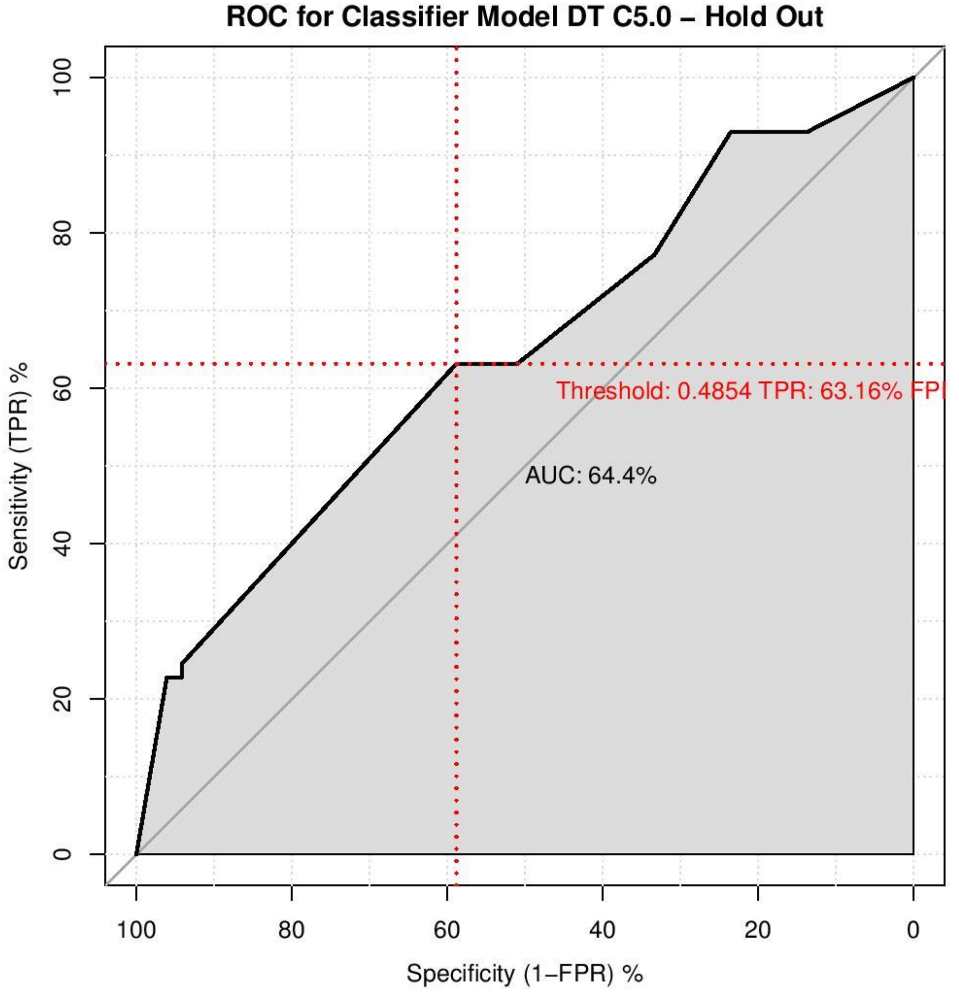
The following are the strength of the variable in the holdout decision tree:



	Strength
breaches	98.61
aRespondentC_SternB	90.74
Economicsector_TertiaryDElectricity	36.81
Economicsector_PrimaryBMining	35.42
Arbitralrules_UNCITRAL	24.07
RespondentHDI	20.83
Economicsector_TertiaryFConstruction	18.29
Arbitralrules_ICSID	15.74
Arbitralrules_ICSIDAF	13.43
aRespondentC_LandauT	13.43
ClaimantHDI	13.43
aClaimantC_BrowerCN	12.73
aRespondentC_ThomasJC	12.73
aClaimantC_HanotiauB	12.04
aRespondentC_DouglasZ	12.04
aPresidentC_FortierLY	3.94
aClaimantC_FortierLY	2.08
aPresidentC_ArmestoJ	0.69
Arbitralrules_ICC	0.00
Arbitralrules_SCC	0.00
aPresidentC_KaufmannKohlerG	0.00
aPresidentC_VeederVV	0.00
aPresidentC_TercierP	0.00
aClaimantC_AlexandrovSA	0.00
aClaimantC_OrregoVic	0.00
aClaimantC_Grigera	0.00
aClaimantC_BeecheyJ	0.00
aRespondentC_SandsP	0.00

The significance of rules derived can be seen in the graph below.


DT C5.0 - Hold Out
80
60
40
20
0
breaches
ondentC_SternB
tiaryDElectricity
_PrimaryBMining
ules_UNCITRAL
RespondentHDI
ryFConstruction
bitralrules_ICSID
alrules_ICSIDAF
dentC_LandauT
ClaimantHDI
antC_BrowerCN
ntC_ThomasJC
antC_HanotiauB
dentC_DouglasZ
dentC_FortierLY
nantC_FortierLY
dentC_ArmestoJ
rbitralrules_ICC
bitralrules_SCC
aufmannKohlerG
entC_VeederVV
identC_TercierP
_AlexandrovSA
antC_OrregoVic
imantC_Grigera
antC_BeecheyJ
ndentC_SandsP

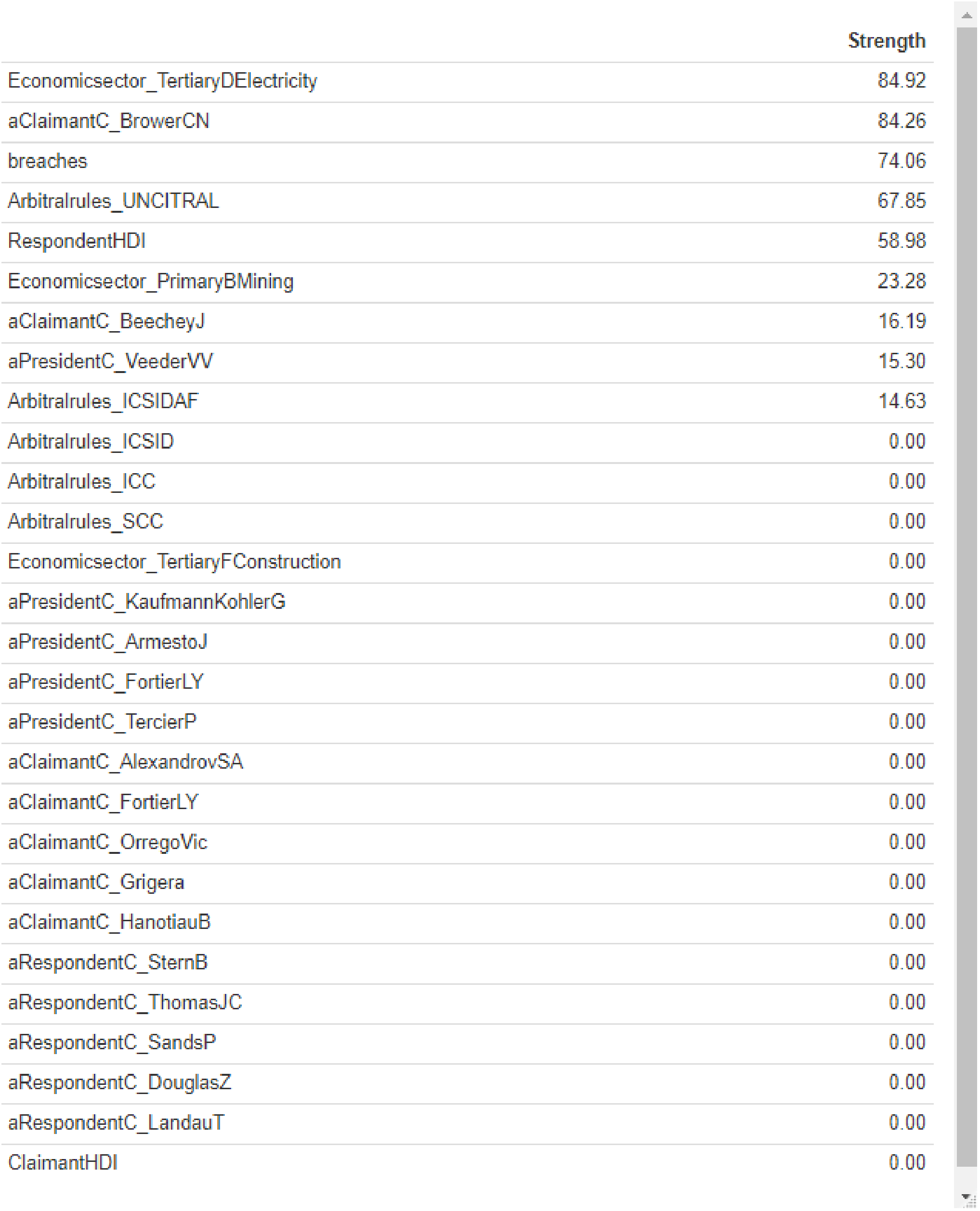

ROC for Classifier Model DT C5.0 – Hold Out
Sensitivity (TPR) %
0
20
40
60
80
100
Specificity (1–FPR) %
100
80
60
40
20
0
Threshold: 0.4854 TPR: 63.16% FP
AUC: 64.4%

5.2.2 Decision Tree K-fold

For the K-fold decision tree 6 folds were set. 12 rules were derived using the full set of variables. The resulting measurements are:

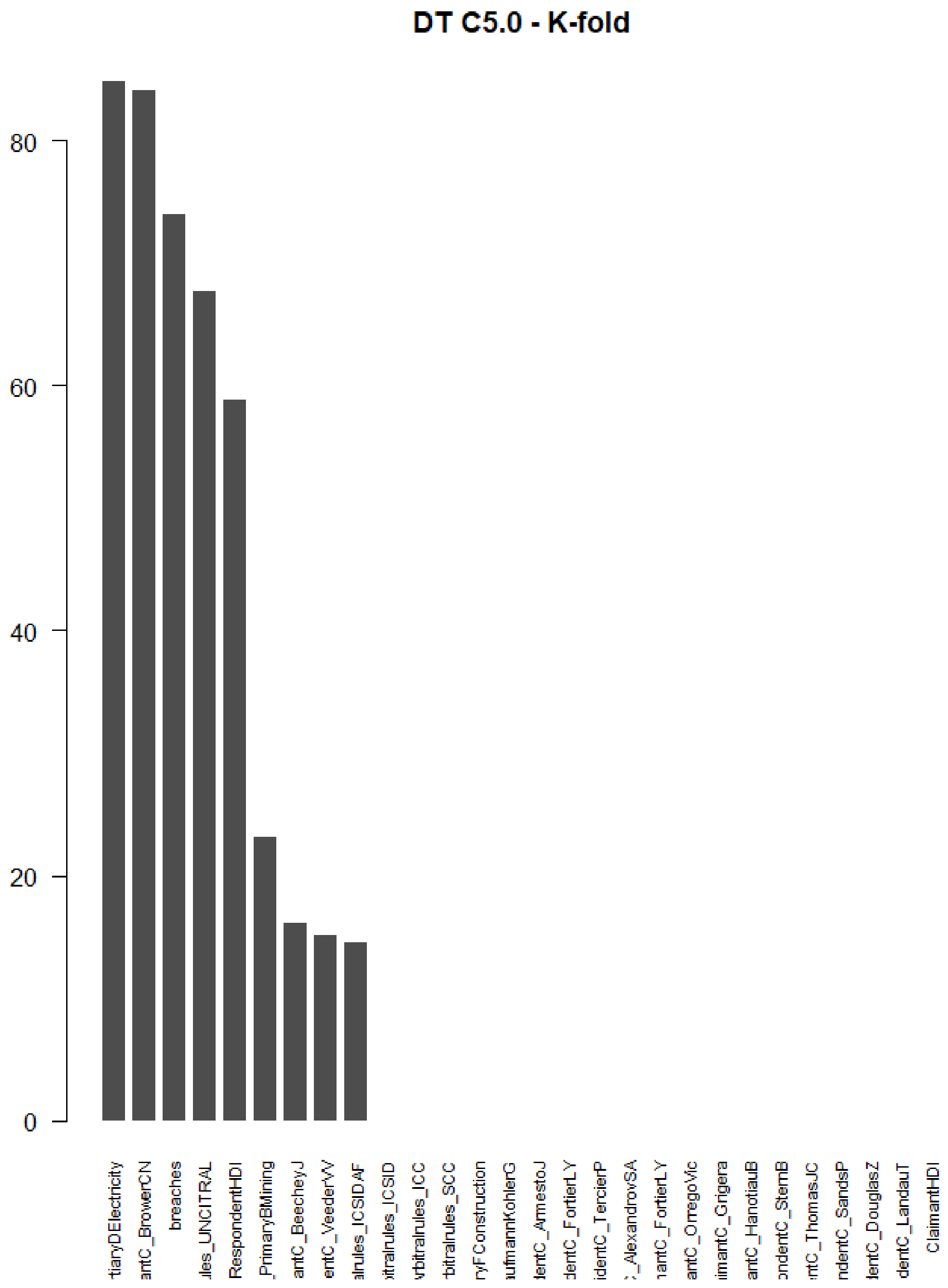
Measure	Averaged Value
TP	26
FN	19
TN	28
FP	14
Accuracy	61.47
pgood	64.55
pbad	58.95
FPR	34.28
TPR	57.36
TNR	65.72
MCC	0.23
Threshold	0.332

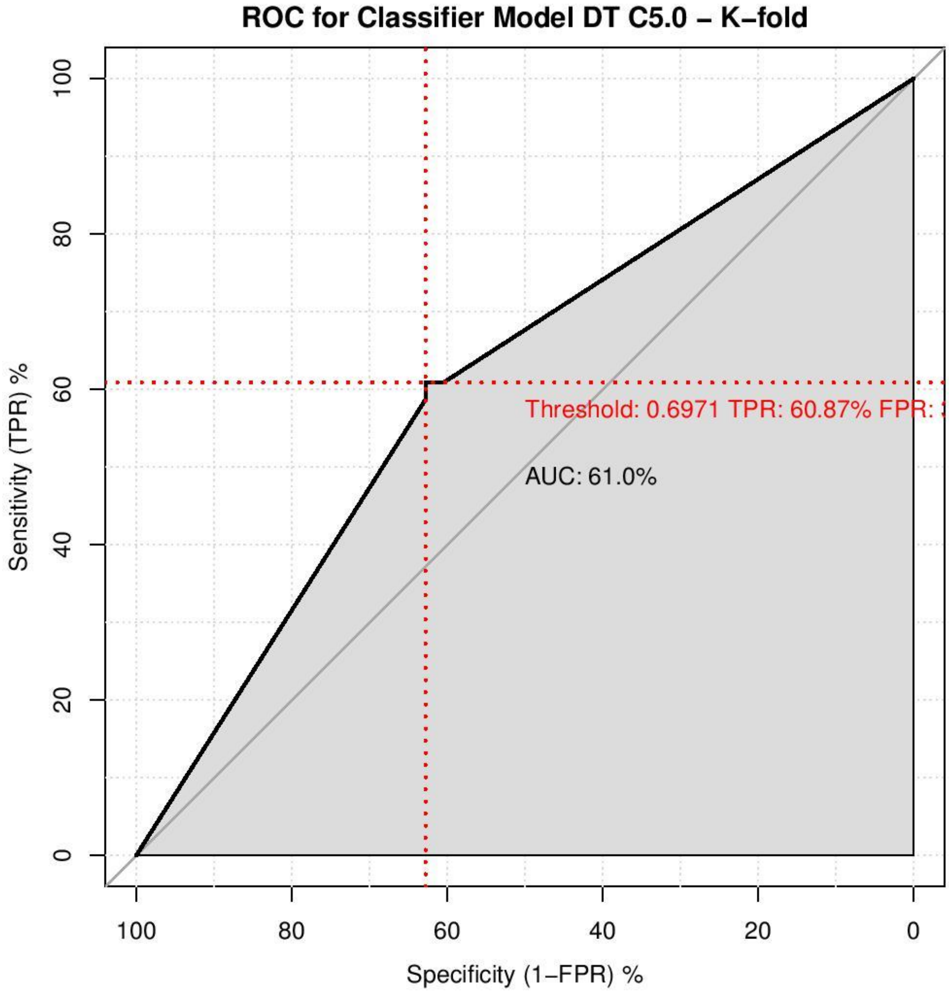
The following are the strength of the variable in the K-fold decision tree:



	Strength
Economicsector_TertiaryDElectricity	84.92
aClaimantC_BrowerCN	84.26
breaches	74.06
Arbitralrules_UNCITRAL	67.85
RespondentHDI	58.98
Economicsector_PrimaryBMining	23.28
aClaimantC_BeecheyJ	16.19
aPresidentC_VeederVV	15.30
Arbitralrules_ICSIDAF	14.63
Arbitralrules_ICSID	0.00
Arbitralrules_ICC	0.00
Arbitralrules_SCC	0.00
Economicsector_TertiaryFConstruction	0.00
aPresidentC_KaufmannKohlerG	0.00
aPresidentC_ArmestoJ	0.00
aPresidentC_FortierLY	0.00
aPresidentC_TercierP	0.00
aClaimantC_AlexandrovSA	0.00
aClaimantC_FortierLY	0.00
aClaimantC_OrregoVic	0.00
aClaimantC_Grigera	0.00
aClaimantC_HanotiauB	0.00
aRespondentC_SternB	0.00
aRespondentC_ThomasJC	0.00
aRespondentC_SandsP	0.00
aRespondentC_DouglasZ	0.00
aRespondentC_LandauT	0.00
ClaimantHDI	0.00

The significance of rules derived can be seen in the graph below.


DT C5.0 - K-fold
80
60
40
20
0
tiaryDElectricity
antC_BrowerCN
breaches
ules_UNCITRAL
RespondentHDI
_PrimaryBMining
antC_BeecheyJ
entC_VeederVV
alrules_ICSIDAF
bitralrules_ICSID
rbitralrules_ICC
bitralrules_SCC
ryFConstruction
aufmannKohlerG
dentC_ArmestoJ
dentC_FortierLY
identC_TercierP
:_AlexandrovSA
nantC_FortierLY
antC_OrregoVic
iimantC_Grigera
antC_HanotiauB
ondentC_SternB
ntC_ThomasJC
ndentC_SandsP
dentC_DouglasZ
dentC_LandauT
ClaimantHDI


ROC for Classifier Model DT C5.0 – K–fold
Sensitivity (TPR) %
Specificity (1–FPR) %
0
20
40
60
80
100
100
80
60
40
20
0
Threshold: 0.6971 TPR: 60.87% FPR:
AUC: 61.0%

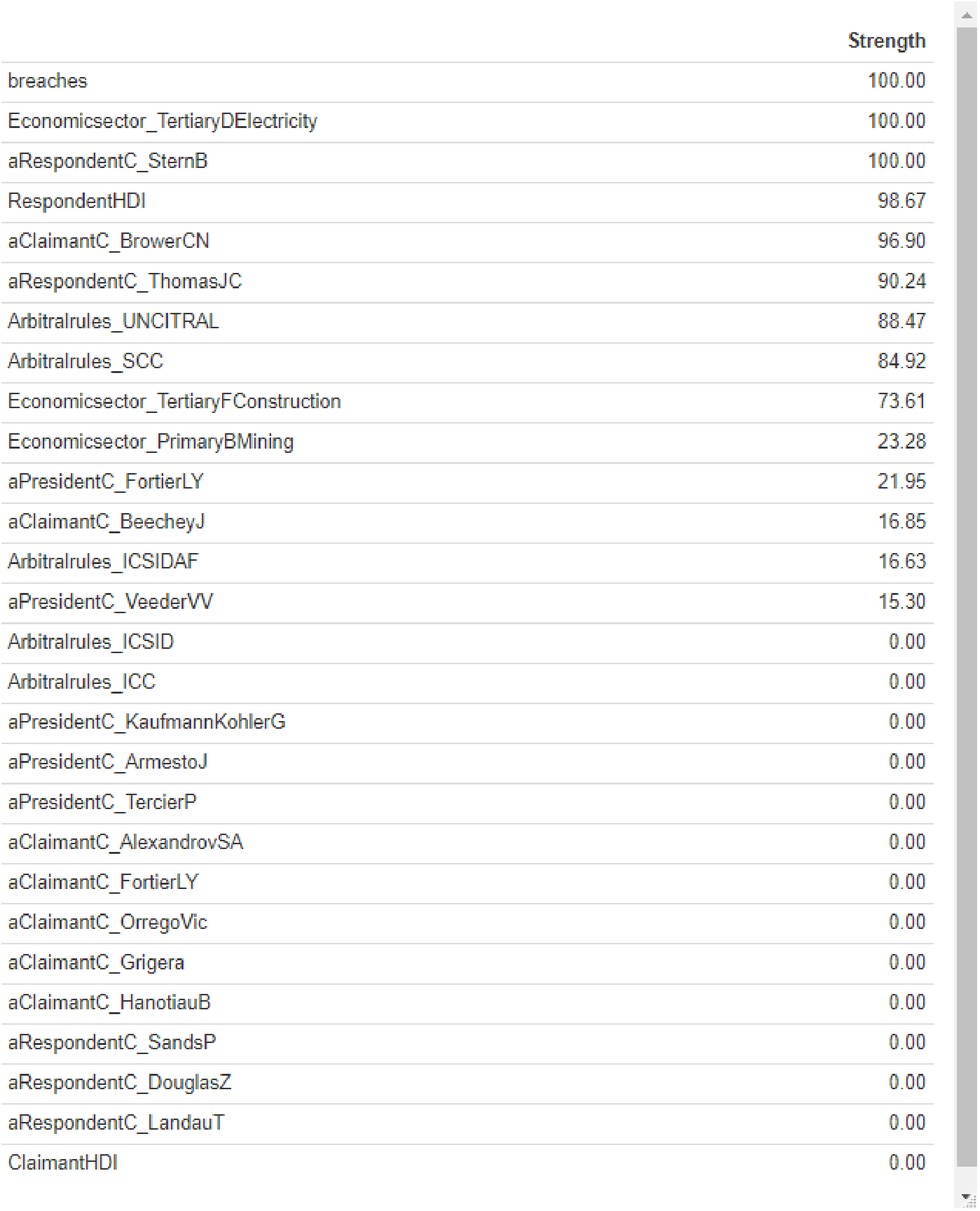
5.2.3 Decision Tree Boosted

A C5.0 decision tree with boost=20 produced the following results. The model was trained using K-fold cross validation and the performance measures were averaged over the 6 folds.

The resulting measurements are:

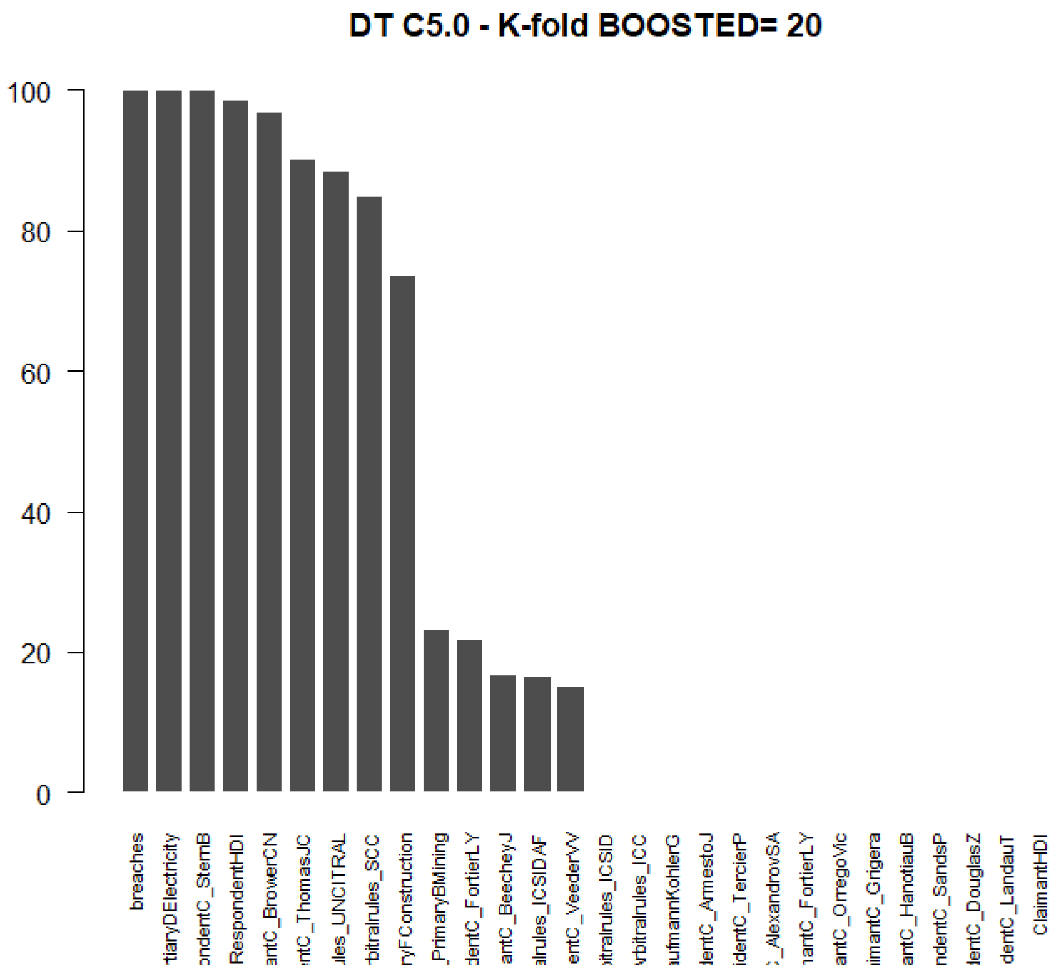
Measure	Averaged Value
TP	25
FN	20
TN	30
FP	13
Accuracy	62.06
pgood	66.39
pbad	59.10
FPR	30.73
TPR	55.38
TNR	69.27
MCC	0.251
Threshold	0.54

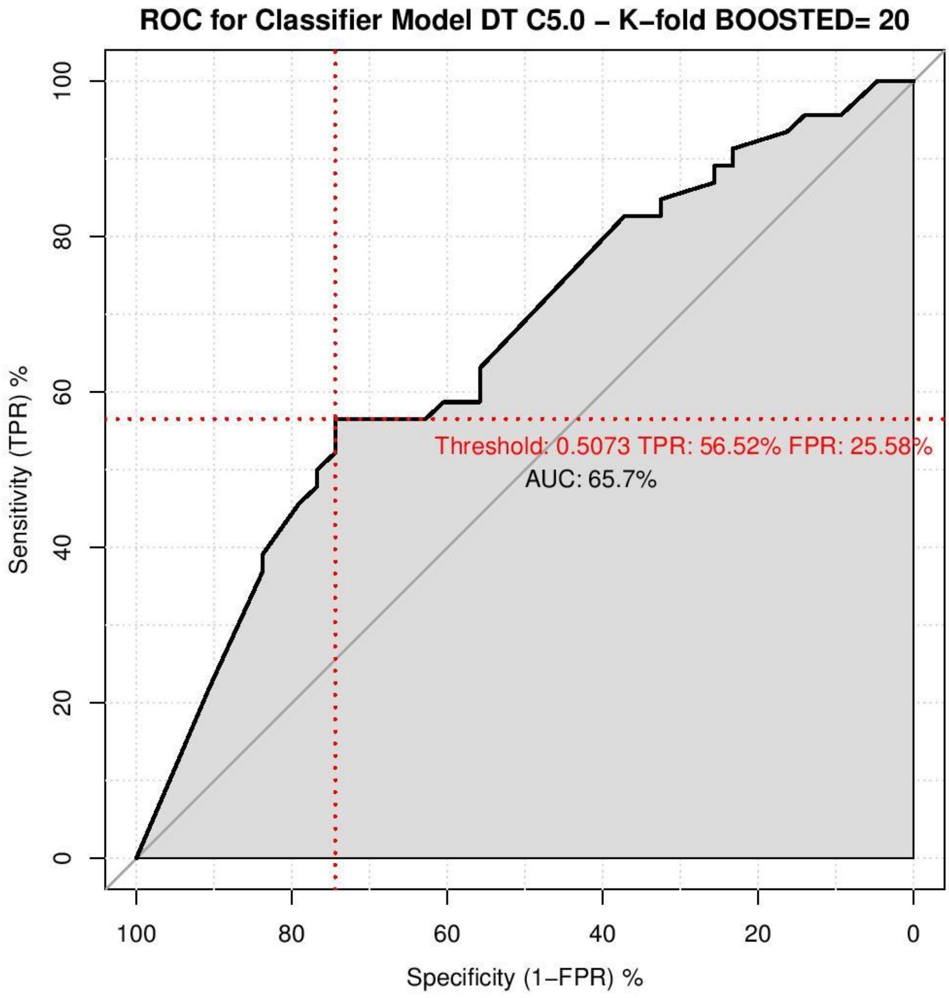
The following are the strength of the variable in the boosted decision tree:



	Strength
breaches	100.00
Economicsector_TertiaryDElectricity	100.00
aRespondentC_SternB	100.00
RespondentHDI	98.67
aClaimantC_BrowerCN	96.90
aRespondentC_ThomasJC	90.24
Arbitralrules_UNCITRAL	88.47
Arbitralrules_SCC	84.92
Economicsector_TertiaryFConstruction	73.61
Economicsector_PrimaryBMining	23.28
aPresidentC_FortierLY	21.95
aClaimantC_BeecheyJ	16.85
Arbitralrules_ICSIDAF	16.63
aPresidentC_VeederVV	15.30
Arbitralrules_ICSID	0.00
Arbitralrules_ICC	0.00
aPresidentC_KaufmannKohlerG	0.00
aPresidentC_ArmestoJ	0.00
aPresidentC_TercierP	0.00
aClaimantC_AlexandrovSA	0.00
aClaimantC_FortierLY	0.00
aClaimantC_OrregoVic	0.00
aClaimantC_Grigera	0.00
aClaimantC_HanotiauB	0.00
aRespondentC_SandsP	0.00
aRespondentC_DouglasZ	0.00
aRespondentC_LandauT	0.00
ClaimantHDI	0.00

The significance of rules derived can be seen in the graph below. In particular, the group noted that 3 features registered 100% importance, indicating that Boosted DT was effectively producing a mini-forest of 3 trees of varying size.


DT C5.0 - K-fold BOOSTED= 20
100
80
60
40
20
0
breaches
tiaryDElectricity
ondentC_SternB
RespondentHDI
antC_BrowerCN
entC_ThomasJC
ules_UNCITRAL
bitralrules_SCC
ryFConstruction
_PrimaryBMining
dentC_FortierLY
antC_BeecheyJ
alrules_ICSIDAF
entC_VeederVV
bitralrules_ICSID
rbitralrules_ICC
aufmannKohlerG
dentC_ArmestoJ
identC_TercierP
_AlexandrovSA
nantC_FortierLY
antC_OrregoVic
imantC_Grigera
antC_HanotiauB
ndentC_SandsP
dentC_DouglasZ
dentC_LandauT
ClaimantHDI


ROC for Classifier Model DT C5.0 – K–fold BOOSTED= 20
Sensitivity (TPR) %
0
20
40
60
80
100
Specificity (1–FPR) %
100
80
60
40
20
0
Threshold: 0.5073 TPR: 56.52% FPR: 25.58%
AUC: 65.7%

5.2.4 Random Forest

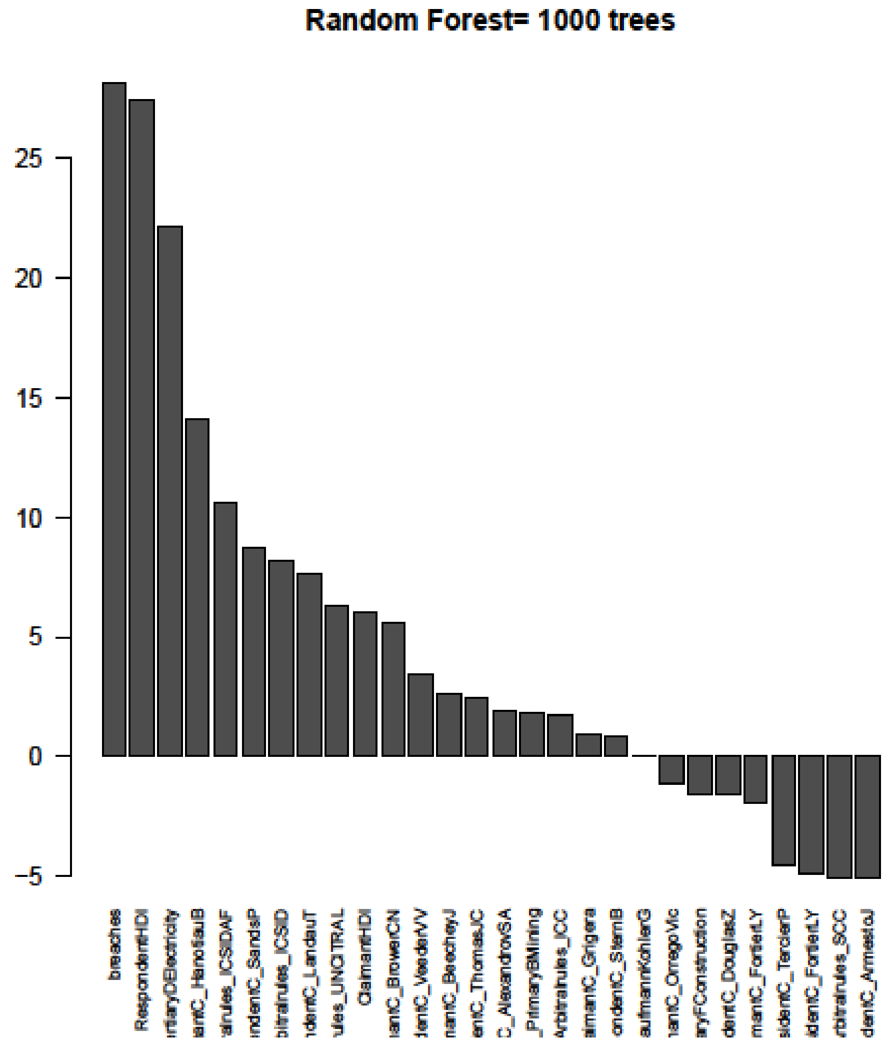
A Random Forest model for classification of the Success class was generated using the standard RandomForest R library.

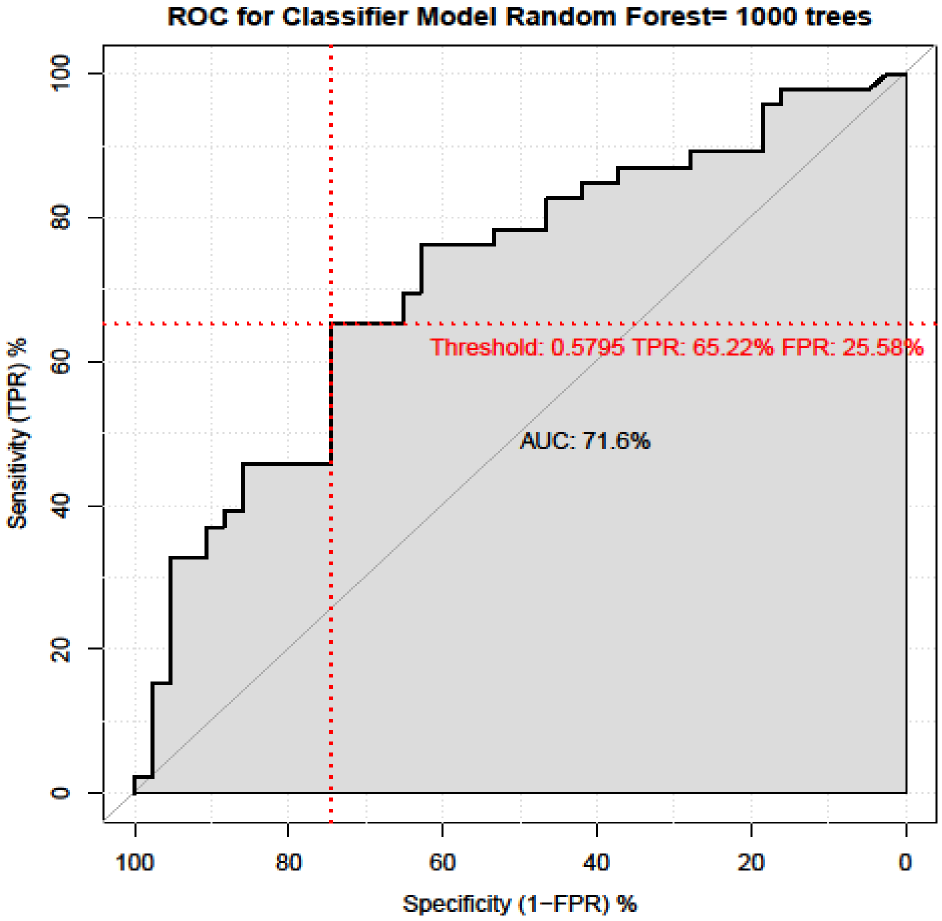
The number of trees was 1000 and the number of candidate variables at each split was the square root of the number of variables, which was 28, giving an mtry=5. The model was trained using K-fold cross validation and the performance measures were averaged over the 6 folds. The threshold was chosen using the minimum Euclidian distance method.

The averaged measures are shown in the table below:

Measure	Averaged Value
TP	27
FN	19
TN	30
FP	12
Accuracy	64.23
FPR	29.22
TPR	58.26
TNR	70.78
MCC	0.293
Threshold	0.538

An example of the variable importance and the ROC chart for one of the folds is shown below.


Random Forest= 1000 trees
25
20
15
10
5
0
−5
breaches
RespondentHDI
rtiaryDElectricity
antC_HanotiauB
rainules_ICSIDAF
ndentC_SandsP
bitrainules_ICSID
ondentC_LandauT
ules_UNCITRAL
ClaimantHDI
iantC_BrowerCN
dentC_VeederVV
mantC_BeecheyJ
entC_ThomasJC
C_AlexandrovSA
_PrimaryBMining
Arbitrainules_ICC
aimantC_Grigera
ondentC_SternB
aufmannKohlerG
iantC_OrregoVic
aryFConstruction
dentC_DouglasZ
mantC_FortierLY
sidentC_TerclerP
identC_FortierLY
rbitrainules_SCC
dentC_ArmestoJ


ROC for Classifier Model Random Forest= 1000 trees
Sensitivity (TPR) %
0
20
40
60
80
100
Specificity (1−FPR) %
100
80
60
40
20
0
Threshold: 0.5795 TPR: 65.22% FPR: 25.58%
AUC: 71.6%

5.2.5 Deep Learning Neural Network

A deep learning neural network model for classification of the Success class was generated using the standard H2O R library with two hidden layers of 5 neurons.

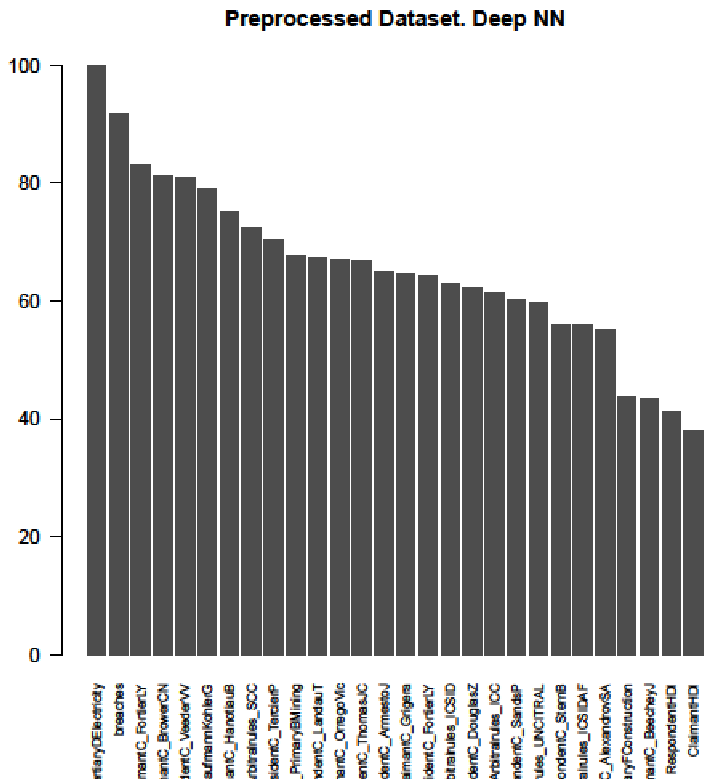
A number of key parameter settings are shown in the table below.

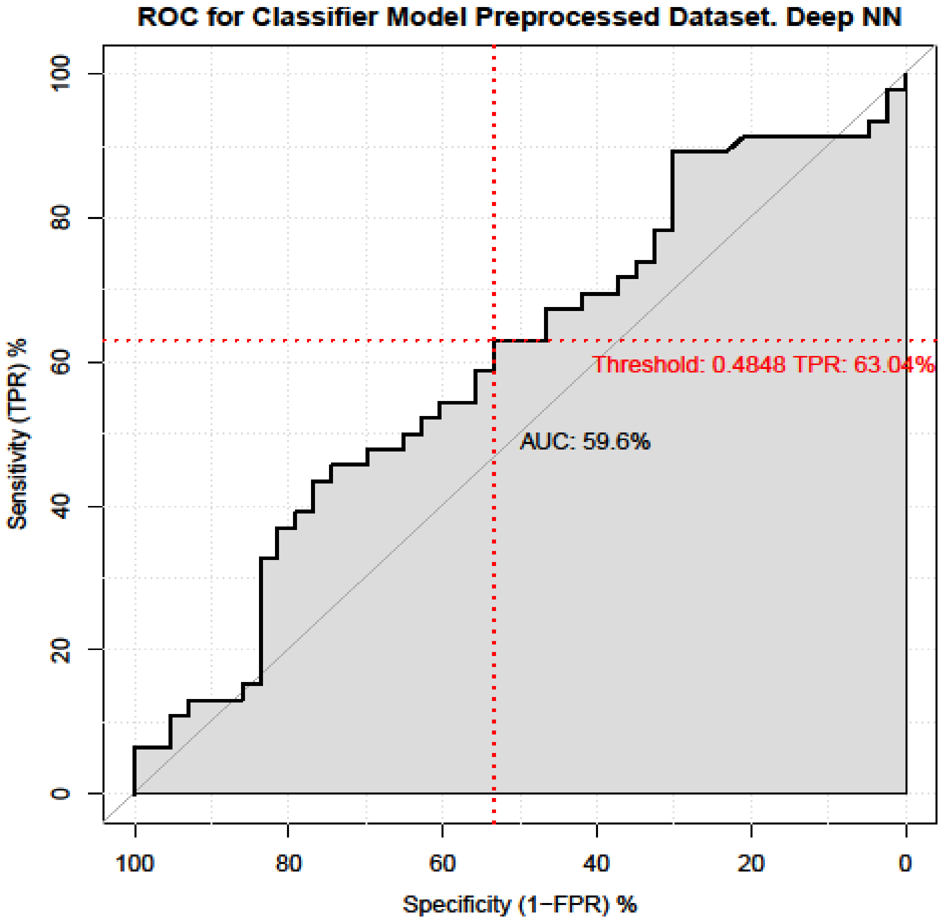
Parameter	Value
Activation	TanhWithDropout
Epochs	100
Neurons	5,5

The averaged measures are shown in the table below.

Measure	Averaged Value
TP	27
FN	19
TN	24
FP	21
Accuracy	57.49
FPR	43.83
TPR	59.28
TNR	56.12
MCC	0.159
Threshold	0.515

An example of the variable importance and the ROC chart for one of the folds is shown below.


Preprocessed Dataset. Deep NN
100
80
60
40
20
0
ıntaryDElectricity
breaches
mantC_FortierLY
ıantC_BrowerCN
ıdentC_VeederW
aufmannKohlerG
antC_HanotiauB
ırbitrarules_SCC
sidentC_TercierP
_PrimaryBMining
ıdentC_LandauT
ıantC_OrregoVic
entC_ThomasJC
dentC_ArmestoJ
aimantC_Grigera
identC_FortierLY
bitrarules_ICSID
dentC_DouglasZ
Arbitrarules_ICC
ındentC_SandsP
ules_UNCITRAL
ondentC_SternB
alrules_ICSIDAF
C_AlexandrovSA
aryFConstruction
ıantC_BeecheyJ
RespondentHDI
ClaimantHDI


ROC for Classifier Model Preprocessed Dataset. Deep NN
Sensitivity (TPR) %
0
20
40
60
80
100
100
80
60
40
20
0
Specificity (1−FPR) %
Threshold: 0.4848 TPR: 63.04%
AUC: 59.6%

5.2.6 Multi-Layer Perceptron

A multi-layer perceptron model for classification of the Success class was generated using the standard H2O R library with a single hidden layer of 10 neurons.

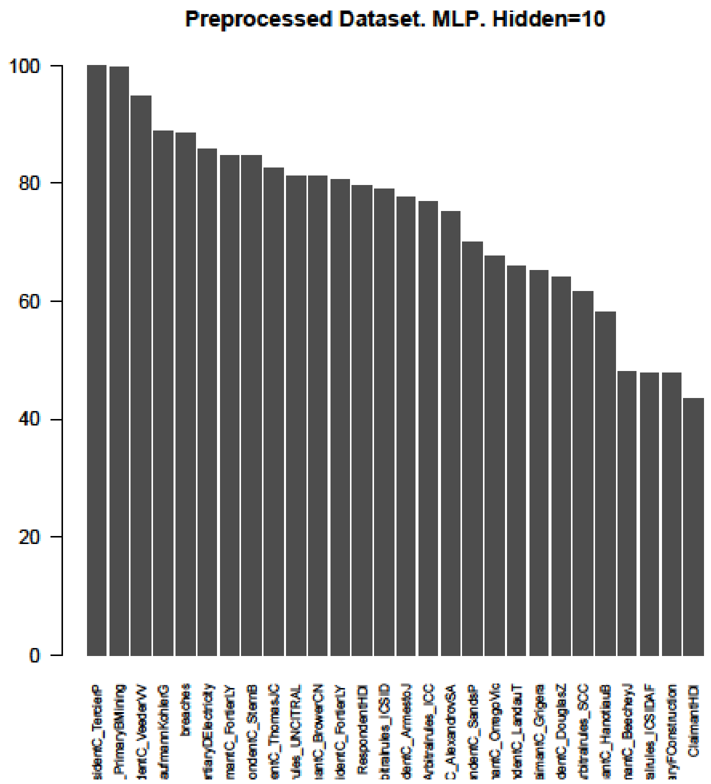
A number of key parameter settings are shown in the table below.

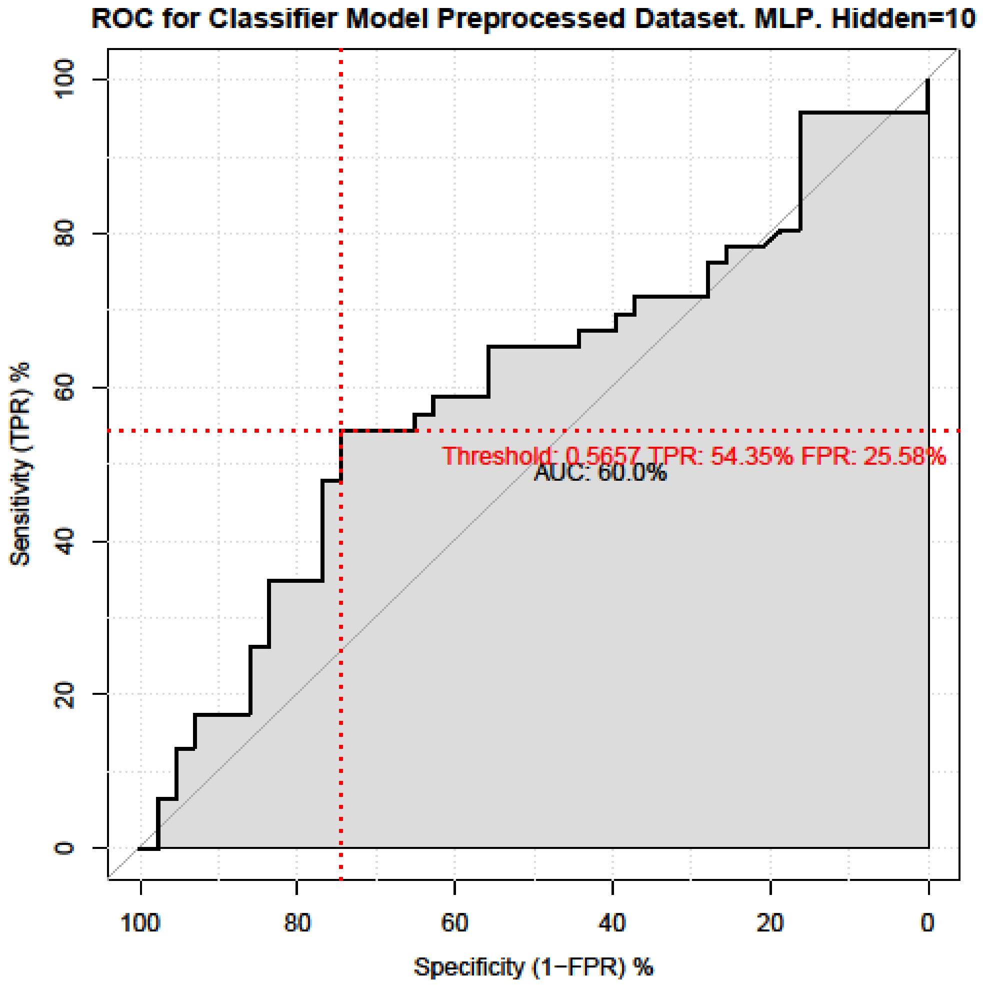
Parameter	Value
Activation	TanhWithDropout
Epochs	100
Neurons	10

The averaged measures are shown in the table below.

Measure	Averaged Value
TP	25
FN	21
TN	29
FP	13
Accuracy	61.12
FPR	31.54
TPR	54.29
TNR	68.46
MCC	0.23
Threshold	0.507

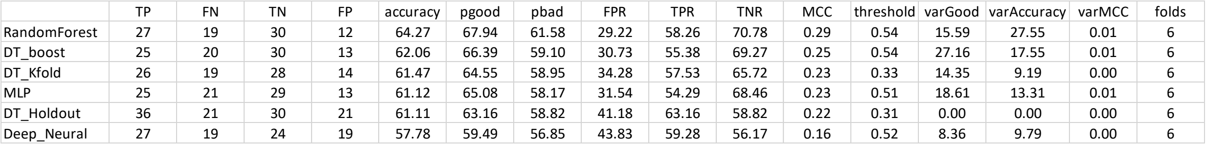
An example of the variable importance and the ROC chart for one of the folds is shown below.


Preprocessed Dataset. MLP. Hidden=10
100
80
60
40
20
0
sidentC_TercierP
_PrimaryBMining
dentC_VeederVV
aufmannKohlerG
breaches
ntaryDElectricity
nantC_FortierLY
ondentC_SternB
entC_ThomasJC
ules_UNCITRAL
iantC_BrowerCN
identC_FortierLY
RespondentHDI
bitralrules_ICSID
dentC_ArmestoJ
Arbitralrules_ICC
C_AlexandrovSA
indentC_SandsP
nantC_OrregoVic
ndentC_LandauT
aimantC_Grigera
dentC_DouglasZ
rbitralrules_SCC
antC_HanotiauB
nantC_BeecheyJ
alrules_ICSIDAF
aryFConstruction
ClaimantHDI


ROC for Classifier Model Preprocessed Dataset. MLP. Hidden=10
Sensitivity (TPR) %
0
20
40
60
80
100
Specificity (1−FPR) %
100
80
60
40
20
0
Threshold: 0.5657 TPR: 54.35% FPR: 25.58%
AUC: 60.0%

5.2.7 Model Assessment including Summary of Results

The following table was produced summarising the averaged results across the six models used to predict case outcome on the mixed continuous and categorical data:



	TP	FN	TN	FP	accuracy	pgood	pbad	FPR	TPR	TNR	MCC	threshold	varGood	varAccuracy	varMCC	folds
RandomForest	27	19	30	12	64.27	67.94	61.58	29.22	58.26	70.78	0.29	0.54	15.59	27.55	0.01	6
DT_boost	25	20	30	13	62.06	66.39	59.10	30.73	55.38	69.27	0.25	0.54	27.16	17.55	0.01	6
DT_Kfold	26	19	28	14	61.47	64.55	58.95	34.28	57.53	65.72	0.23	0.33	14.35	9.19	0.00	6
MLP	25	21	29	13	61.12	65.08	58.17	31.54	54.29	68.46	0.23	0.51	18.61	13.31	0.01	6
DT_Holdout	36	21	30	21	61.11	63.16	58.82	41.18	63.16	58.82	0.22	0.31	0.00	0.00	0.00	6
Deep_Neural	27	19	24	19	57.78	59.49	56.85	43.83	59.28	56.17	0.16	0.52	8.36	9.79	0.00	6

The key result observed from this modelling phase was that in the generation of the confusion matrix and related statistical measures, the Random Forest and DT (Boosted) models performed the best on most measures. Specifically, they each had an MCC in excess of 0.25, and the highest accuracy rate (greater than 0.62). They also had the greatest AUC in the ROC curves indicating superior performance in TPR vs FPR. Only in variance did Random Forest and DT Boosted (slightly) underperform relative to Deep Neural Network and MLP.

Although Random Forest and DT Boosted had an MCC of only 0.29/0.25 indicating mild correlation between overall class prediction and actual class, the group noted that prediction of Successful cases outperformed prediction of Failure cases across all models. Most importantly Random Forest and DT Boosted had the highest win-rate (approximately 2 in 3 - the p-good measure) i.e. each successfully predicted winning cases approx. 66% of the time in the test samples of 90 cases. This interesting result compares favourably with the overall sample rate of 51.8% and is discussed below in the Evaluation section.

Additionally, the DT Boosted model provided greater explanation relative to the NN/MLP approaches, in terms of the importance of key variables/rules in predicting outcomes. Specifically, DT Boosted indicates that a mere 5 variables are most important, plus a further 4 more minor variables. This compares with NN/MLP which indicate a much more even weight across the entire variable field (28). Evidence of a similar effect is also seen for Random Forest where 3 variables (Breach; Respondent HDI; Sector – Electricity) are more important than the rest (followed by a further 6 variables), and which same 3 lead variables are found in the set of first 5 variables of DT Boosted.

Overall, it is suggested that DT Boosted performed best in terms of prediction of outcome with limited variance (mean 0.66, SD = 0.052), combined with ability to explain the results. That said, it would seem sensible to combine DT Boosted with Random Forest in future modelling exercises.

6 Section 6 – Evaluation and Future Deployment

In summary, the group believes the project has been a qualified success, recognising that notwithstanding the small sample size each of the data pre-processing steps took far longer than anticipated. In terms of meeting the goals that were set, the project has been able to test in part each of the hypotheses raised in Section 2 – Problem Definition. However, given limited time and resources available, the group is also conscious that the hypotheses could be tackled in the future in greater detail. Suggestions for further work are identified below, which are encouraged now that the time-consuming data preparation work has been completed.

6.1 Evaluation of the Linear Regression Models

As noted earlier the Linear Regression Models used were very simple but good enough to assess additional hypotheses 1, 2 and 5. The Linear regression models clearly showed no meaningful correlation between HDI (respondent or Claimant) and amounts awarded, nor between HDI and damages ratio. But they did reveal significant correlation between claimed and awarded amount. Accordingly, the group was confident in (partially) rejecting/accepting the hypotheses as follows:

- Respondent States with a lower HDI ranking are penalised by arbitrators disproportionately when compared to higher ranking States;

  Hypothesis ***rejected*** in so far as there was no evidence that lower HDI Respondents are penalised with (relatively) larger amounts awarded against them, all other things being equal. It is however, evident that Respondent States subject to Damages Awards are likely to have an HDI ranking lower than that of the Claimant State. This should be expected given that higher HDI states are more likely to be capital exporting states, and which capital is likely to be exported to lower HDI states;

- Claimants from higher ranking HDI States receive preferential treatment;

  Hypothesis ***rejected*** in so far as there was no evidence that higher HDI Claimants are rewarded with (relatively) larger amounts awarded to them, all other things being equal;

- Damages claimed is positively correlated to damages awarded;
  Hypothesis accepted in so far as there was some evidence (R-squared 0.47) that in successful cases damages awarded is positively correlated to damages claimed and in fact that the best estimate of the Damages Ratio is 0.34 plus/minus the residual error term. In other words, “if you don’t ask, you don’t get”. But it was also acknowledged that the R-squared value implies significant residual variance unexplained by the regression model.

6.2 Evaluation of the Classifier Models

As noted above, the group explored whether the core and certain of the additional hypotheses (numbers 3, 4 and 6) were supported by the model output on the 540 concluded cases. Again, it was noted that the sample of size of 540 cases was small, and so caution is necessary when drawing conclusions. However, based on the results from the DT Boosted and Random Forest Models, the group was confident in partially rejecting/accepting the hypotheses as follows:

- ***Core hypothesis***: individual factual case analysis should be fundamental and determinative when deciding on the likely outcome of an investment treaty case. Or can it be said that the UNCTAD model appears to explain (some of) the variance in case outcomes?

  Core hypothesis partly rejected in so far as there was evidence that the DT Boosted and Random Forest models (among others) with a focus on a limited number of exogenous variables appeared able to predict winning cases with approximately 66% probability versus sample mean of 51.8%, while noting that overall accuracy and prediction of losing cases was less impressive.

  The group analysed the P-good results of DT Boosted and Random Forest further to investigate their significance. In general terms, a lawyer with a success rate in winning 66% of his or her cases might be considered good in the context of a field where the mean is 51.8%. In fact, few lawyers would profess to do any better.

In terms of statistical significance, the group considered this result in the context of a sample test size of 88 cases, for which the mean actual success rate was 0.518. Assuming this to be a binomial distribution would give a sample standard deviation of 0.053 about this mean. Approximating the sample binomial distribution with a normal distribution, enabled the group to make an estimate of the significance of this result. Specifically, the probability of getting a sample score of 0.66 success rate in a random sample of 88 cases was less than 1% leading the group to conclude that the evidence against the core hypothesis was significant at the 0.01 level. Or put another way there was some evidence that the model had predictive power as to case outcomes, whilst recognising that 66% falls a long way short of 100% success rate.

The group was reluctant to get too carried away with this result given the material variance in p-good outcomes. Nonetheless, the model provides support that a review of certain identified simple exogenous factors (most important variables) should be made as part of making case assessments, and that this should be cost effective given the ease with which the features could be identified.

- Claims in heavy industry sectors (Mining, Power, Construction) outperform other cases;
  Core hypothesis partly accepted in so far as there was evidence that the DT Boosted and Random Forest models assigned significant importance to this variable, and in doing so were able to significantly improve p-good. In particular, Power-Electricity appeared as a very strong variable in both models in predicting winning cases. In other words, although we did not reach a conclusion as to whether these cases outperform in isolation (simple statistical measure), we did obtain evidence that using the rules generated by the model on these variables improves overall win-rate;

- Direct expropriation cases perform better than other cases brought under the lower treaty breach standards;
  Core hypothesis partly accepted in so far as there was evidence that the DT Boosted and Random Forest models assigned significant importance to this variable, and in doing so were able to significantly improve p-good. In particular, Direct Expropriation appeared as a very strong variable in both models in predicting winning cases. Again,

although we did not reach a conclusion as to whether these cases outperform in isolation (simple statistical measure), we did obtain evidence that using the rules generated by the model on these variables improves overall win-rate;

- Choice of (some of) the "usual suspect" Arbitrators by Claimant, Respondent or as President (versus choosing another) can have an impact on case outcome.
  Core hypothesis partly accepted in so far as there was evidence that the DT Boosted and Random Forest models assigned significant importance to this variable, and in doing so were able to significantly improve p-good. Various different arbitrators appeared as a very strong variable in both models in predicting winning cases. Again, although we did not reach a conclusion as to which arbitrators generate under or overperformance in isolation, we did obtain evidence that using the rules generated by the model on these variables improves overall win-rate.

6.3 Suggestions for Future Projects / Deployment of Model

- Subject to availability of resources, a natural next step would be to deploy the various models against the pending cases (341 cases) and then observe performance over the course of the coming year (2020) when it might be expected that 50-70 of the pending cases will have concluded.
- Additional work could be done on the variables considered so far to isolate the impact specific features have on case outcome, e.g. does a specific arbitrator or combination of arbitrators make a difference? Or perhaps a specific combination of 2 or more variables? A more formalised ranking of the features could be undertaken to measure relative importance.
- Lastly, a natural extension of the model would be to combine it with a factual enquiry into case sample, reflecting additional factual features proposed by domain experts. Potentially these features could be extracted by NLP techniques from legal case documents. The hypothesis that would then fall to be tested would be that these additional factual features could make a significant improvement in p-good over the 540 sample of concluded cases. Ultimately the hope would be that such a model combining easily already observed key features plus domain expert identified factual features could be utilised successfully for future case outcome prediction.