International Journal of Control and Automation
Vol. 13, No. 4, (2020), pp. 1312-1325

Unboxing the Classification for Visualization of the Outcomes with Naïve User Perspective

[1, *]Pawan Kumar and [2]Manmohan Sharma
[1,2]Lovely Professional University, Phagwara, India
pawan.11522@lpu.co.in, manmohan.21909@lpu.co.in

Abstract

Most of the ML-based models behave like a black-box in the sense that their behaviour is not easily understandable to naïve users. This paper proposes a two-layer framework for evaluation of learning acquired by an ML model and facilitate trust of a human user through on-demand explanations. To verify the reliability of a model learning, the idea is to understand its behaviour and assessing to what extent the model has incorporated important characteristics of the provided dataset. Information gain measures using entropy and Gini index are used to compute dataset characteristics. Feature importance, global surrogate model, and local surrogate models are used to understand model behaviour. Measuring the degree of agreement between the provided dataset and the learned model is modelled as a 2-Judge and n-participants rank correlation problem. A positive association using Spearman's rank correlation acts as an indicator of the reliability of the learned ML-based model.

Keywords: machine learning, classification, human interpretability, information gain, feature importance, interpretable machine learning.

1. Introduction

Machine Learning (ML) refers to algorithms that can make a machine learn from examples [1]. Software engineering has opened its arena and software designers are coming up with tools with ML capabilities integrated into them. These tools along with the availability of secondary datasets on different portals like Kaggle and UCI repository have made the life of researchers easier. Using these tools, even the users who are not ML experts can use ML in their respective problem domains. As we are gearing up to take ML applications to the end-user experience, there is a rising need of giving justifications behind outcomes of an ML-based model. Although some of the ML-based models like linear regression and decision trees are inherently interpretable, most of the advanced ML algorithms like Random Forests and Neural networks are not. These advanced ML models lack in justifying their outcome in a manner understandable to the layman, despite being extremely good in classification accuracy. This lacking, termed as lack of human interpretability of outcomes, is a hurdle in improving penetration of ML usage in fields involving critical decision-making like supporting human experts in medical diagnosis. As another example, companies like Amazon have started giving explanations of why a particular product is being recommended to a user? ML models are working behind to analyze the navigation and shopping behaviour of customers visiting online shopping portals. Conferring human interpretability to ML models offers advantages like the discovery of new knowledge, model debugging, ensuring fairness, and facilitating trust.

During recent years, the ML community has shown increased interest in the field of human interpretability. The existing work can be categorized into model-specific and model-agnostic

ISSN: 2005-4297 IJCA

methods. Model-specific methods apply to a particular underlying ML algorithm whereas model-agnostic refer to techniques that apply to all ML-based models irrespective of the underlying ML model.

Model-specific approaches: A framework named 'inTrees' that consists of algorithms to extract, process, prune and summarize rules from a tree ensemble has been proposed [2]. Use of two models, 'P' for prediction and 'I' for interpreting has been proposed to approximate the learned complex tree ensemble by a simple interpretable model using KL-divergence as proximity measure [3]. An approach benefitting from two unique aspects of tree ensembles i.e. leveraging tree structure and naturally-learned similarity measure has been proposed for interpreting tree ensembles by finding prototypes in tree space [4]. GENESIM algorithm for extraction of a single interpretable model from an initial population of decision trees using a genetic algorithm has been proposed [5]. Deep Taylor decomposition technique has been used to decompose decisions of neural networks in terms of contribution from input elements [6]. Layer-wise relevance propagation (LRP) framework has been proposed to explain predictions of a deep neural network [7].

Model-agnostic approaches: A general method to explain a prediction outcome in terms of individual contributions of features using concepts of coalitional game theory has been proposed [8]. Local explanation vector that is an estimation of local gradients has been used to understand instance-specific outcome [9]. Algorithm LIME has been proposed to explain the prediction outcome of any classifier by learning an interpretable model locally around the prediction [10]. Use of input gradient i.e. partial derivative of the model with respect to the input is proposed for interpreting any model [11]. The paper [12] advocates the use of model-agnostic approach covering advantages and associated challenges. Using programs as model-agnostic local explanations have also been proposed [13].

Contribution: For a successful ML-based solution, both, the reliability of the model learning as well as its ability to explain prediction outcomes are important. The contribution of this research work is to propose a framework that can facilitate the evaluation of an ML-based solution in terms of reliability of the learning process and facilitating trust of the human user. The proposed framework has a two-layer architecture. The first layer focuses on the reliability of the learning happened. The intuitive idea used is that reliable learning by an ML model should always result in incorporating important dataset characteristics into its reasoning. The work in this layer aims to answer the following research question: Given a dataset and a black-box ML model learned from that dataset, can dataset measures like entropy and Gini index along with model-agnostic interpretability techniques be used to verify the reliability of the learning process by measuring the degree of agreement between dataset characteristics and model behaviour? The second layer focuses on facilitating trust in the ML model through unboxing the classification outcomes using global and local surrogate models.

The proposed approach was evaluated using two datasets, a standard dataset and a primary dataset. A positive agreement between important dataset characteristics and ML model behaviour was observed for both the datasets, indicating the reliability of the learning process. Also, a comparison of the degree of this agreement for the two datasets was in sync with the quality of learning happened in the respective ML models. On-demand explanations of classification outcomes facilitate trust of human users in the ML-based solution.

ISSN: 2005-4297 IJCA

Organization: Section 2 describes the material (subjects) used, ML algorithms used for classification, interpretability techniques, and experiments designed. Section 3 describes our mathematical framework and the pseudo-code of the exact procedure. Section 4 compiles the results and observations of the experiments conducted. Section 5 concludes the findings of this research and possible future directions.

2. Methods

This section describes how the data was collected and analyzed.

2.1 Materials (Dataset) used

To evaluate the proposed approach, the following two datasets, one primary and one secondary were used:

(i) Telco customer-churn (tcc): 'tcc' is a standard dataset available on Kaggle, consisting of customer records. The target variable is 'Churn' with values as 'Yes' or 'No'. The dataset consisted of 7032 observations and 21variables. There were 1869 customers with churn status 'Yes' and 5163 customers with churn status 'No', giving a baseline accuracy of 0.734. Baseline accuracy is the ratio of majority class frequency to total observations.

(ii) **Freshmen students (freshmen)**: 'freshmen' is a primary dataset, consisting of students who took admission in an educational institute in North India, during 2018 admission year. The target variable was 'JoiningStatus' with values as 'Joined' or 'Lost'. The dataset consisted of 13125 observations and 25 variables. Out of 13125 students, 8374 joined while remaining 4751 did not join, giving a baseline accuracy of 0.638. Table 1 shows the structure of 'freshmen' dataset with the type and description of each attribute.

Table 1. Structure of the 'freshmen' dataset

Attribute	DataType	Description
RegistrationNumber	Int	8-digit Unique ID for each student
AdmissionMonth	Int	Month of admission e.g 5 = May, 6 = June
Gender	Factor	Student gender, F - Female, M – Male
State	Factor	Home state of student
HomeTownType	Factor	Rural, Urban (Metropolitan), Urban (Town)
BatchYear	Factor	Admission year i.e. 2017 or 2018
ProgramName	Factor	BBA, MBA, B. Arch. etc.
Discipline	Factor	Agriculture, Management etc.
QualifyingExam	Factor	Eligibility qualification e.g. 10+2, Graduation
MarksPercent	Factor	Marks in the qualifying exam
CategoryCode	Factor	General, SC, ST etc.
TransportAvailed	Factor	Transport facility of university availed?
LoanLetter	Factor	Education loan availed? [Yes/No]
PreviouslyStudied	Factor	Whether the student studied earlier? [Yes/No]
HostelAvailed	Factor	Hostel availed or not? [Yes/No]
MessAvailed	Factor	Mess availed or not? [Yes/No]
ScholarshipPercentag	Numeric	Scholarship amount as a percentage of tuition
ScholarshipBracket	Factor	High, Low, Medium
EconomicCondition	Factor	AboveAverage, Average. BelowAverage,
FeePaidPercentage	Numeric	Percentage of the fee paid so far by student

ISSN: 2005-4297 IJCA

MediumOfStudy	Factor	English, NonEnglish
FeePaidCategorized	Factor	High, Low, Medium
MarksCategory	Factor	Excellent, Fail, FirstDivision, SecondDivison,
HostelOrTransport	Factor	Hostel or transport availed? [Yes/No]
StudentStatus	Factor	Joined, Lost (Did not join)

2.2 ML algorithms and techniques used

2.2.1 Learning and evaluating ML model

Popular classification algorithms, namely, LR, NB, CART and RF algorithms were explored to learn an accurate ML model. LR and NB estimate the probabilities of each class of target variable using the logistic function and Bayes theorem, respectively. CART is an interpretable decision tree algorithm [14]. RF is an ensemble approach and work by creating a forest of trees and using majority voting [15]. Classification accuracy(acc) and F1-score are used as performance metrics.

2.2.2 Measuring dataset characteristics

Information gain measures using entropy and Gini index were used. Information gain measures the importance of an attribute. Entropy is used to measure information gain if a particular feature is selected for splitting [16]. Gini index is a measure of inequality in distribution [17] and is always in the range of 0 to 1.

2.2.3 Understanding the behaviour of the ML model

Interpretability of model outcomes is of important concern in many decision critical applications [18,19]. Model-agnostic [20] interpretability techniques were used to understand model behaviour.

Feature importance: It is a measure of the increase in the model's error rate when the values of a feature are permuted. A higher increase indicates higher importance for the feature. Feature importance provides a global insight into model behaviour. The measure of Error rate = 1- AUC, was used.

Global surrogate model: It is an approximation of a complex black box model using a simpler model that may not be that accurate but is easy to interpret. R squared value is taken as a measure of how well our surrogate model replicates the original black-box model. It measures the amount of variance captured by our surrogate model.

Local surrogate model: To explain the outcome for an individual instance, LIME (local interpretable model-agnostic explanations) based explanations were used. It shows the most contributing features towards the outcome for that instance. The learned model should be a good approximation of the black box model locally, but it does not have to be so globally.

2.3 Experiments conducted

The experiments designed for each of the two datasets are described in Table 2 along with the underlying motivation.

ISSN: 2005-4297 IJCA

Table 2. Experiments conducted

Experiment #	Experiment Description	Underlying Motivation
Building an ML model	Exploring LR, NB, CART and RF algorithms to build an accurate machine learning model	Selection of an accurate ML model
Compute dataset characteristics	Identifying important features from the dataset using information gain measures	What are important features as per dataset?
Diagnose model behaviour	Understanding model behaviour using model-agnostic interpretability techniques	What reasoning model is using for giving outcomes?
Agreement between model and dataset	Modelling as a 2-Judge n-participant rank correlation problem	Assessing the extent to which model has incorporated dataset characteristics
Unboxing of classification outcomes	Explaining classification outcomes using local and global surrogate models	Facilitating trust in the ML-based solution

3. The proposed framework for reliability and trust

Table 3 describes the acronyms used in the proposed mathematical framework.

Table 3. Acronyms description

Acronym	Description
$IG_{\{entropy,f\}}$	IG for feature 'f' using entropy
$IG_{\{gini,f\}}$	IG for feature 'f' using gini
f_{Imp}	Importance measure for feature 'f' as per model
p(x)	Probability of class 'x' for an instance
$R_{\{entropy,f\}}$	Rank assigned to a feature 'f' using $IG_{\{entropy,f\}}$
$R_{\{gini,f\}}$	Rank assigned to a feature 'f' using $IG_{\{gini,f\}}$
$R_{\{dataset,f\}}$	Average rank to a feature 'f' using $IG_{\{entropy,f\}}$ and
FEATURES	Set of all features used in learning the model
$R_{\{Model,f\}}$	The rank assigned to a feature 'f' based on f_{Imp}

Information gain using entropy for a feature 'f' is computed as:

$$IG_{\{entropy,f\}} = -\sum(p(x)\log 2(p(x)) \quad (A)$$

Information gain using gini for a feature 'f' is computed as:

$$IG_{\{gini,f\}} = 1 - \sum(p(x) * p(x)) \quad (B)$$

Feature importance as per model in terms of increase in error rate is computed as:

$$f_{imp} = 1 - AUC \quad (C)$$

where AUC represents area under the ROC curve.

ISSN: 2005-4297 IJCA

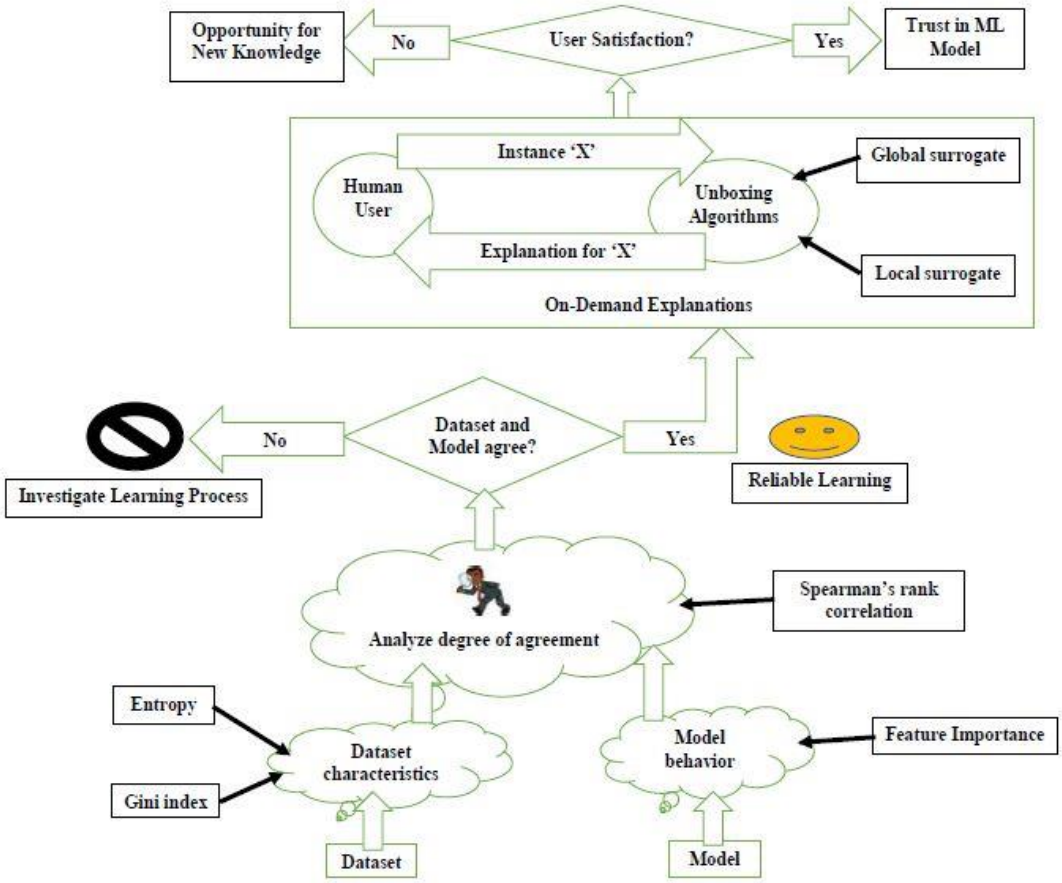

Figure 1. Proposed framework diagram

The input to the system is the provided dataset and a black-box model learned using that dataset. It is assumed that the underlying ML algorithm used for learning the model is unknown. To understand the characteristics of the provided dataset, information gain using entropy and Gini index is computed for each feature. Using these values features are ranked, with rank 1 given to feature with maximum information gain. To find a single rank for each feature, the average of the two ranks is taken. To understand the behaviour of the black-box model, feature importance measure is computed for each feature. Again, each feature is ranked as per its importance to the model with rank 1 assigned to the most important feature. To measure the degree of agreement between the ranks, spearman's rank correlation is used. A positive value of the rank correlation coefficient is taken as a measure of the reliability of the learning process. For winning the trust of human end-user, on-demand explanations are produced. Local and global surrogate models are used for unboxing of classification outcomes. Figure 1, shows the diagrammatic representation of the proposed framework.

The pseudo-code of the exact procedure is depicted below:

ISSN: 2005-4297 IJCA

Pseudo Code: Unboxing a black-box ML model

Assumption: Underlying ML algorithm is not known
FEATURES = Set of all variables used in the learning model
Input: dataset, a black-box ML model
Output: Understanding of the model behaviour
Procedure

1. [Identify important features as per dataset]
   for each 'f' in FEATURES
      Compute $IG_{\{entropy,f\}}$ and $IG_{\{gini,f\}}$ using equations (A) and (B)
   end for
2. [Assign ranks to features]
   for each 'f' in FEATURES
      Assign $R_{\{entropy,f\}}$, using $IG_{\{entropy,f\}}$,
      Assign $R_{\{gini,f\}}$ using $IG_{\{gini,f\}}$,
   end for
3. Compute $R_{\{dataset,f\}}$ as average of $R_{\{entropy,f\}}$ and $R_{\{gini,f\}}$
4. [Identify important features as per Model]
   for each 'f' in FEATURES
      Compute f_{Imp} and $Rank_{\{Model,f\}}$ Using equation (C)
   end for
5. Analyze the degree of agreement between two vectors $R_{\{dataset,f\}}$ and $Rank_{\{Model,f\}}$ using Spearman's rank correlation
6. Facilitate trust by providing on-demand explanations for human users

end procedure

4. Results and discussion

This section compiles the experimental outcomes and discussion on the observations of these experiments.

4.1 Performance evaluation of ML models

Table 4 and 5 compiles the performance metrics of ML models for 'tcc' and 'freshmen' dataset respectively. All the four classification algorithms gave accuracy in the range of 80%, a significant improvement over baseline accuracy of the two datasets, as mentioned in Section 2.1. Moreover, there is a good generalization of classification accuracy from training to test data except in case of RF for customer churn dataset. RF outperformed other models both in terms of accuracy and F1 Score for both the datasets.

Table 4. Performance metrics for 'tcc' dataset

Model	Acc		F1 Score	
	Training	Test	Training	Test
LR	0.795	0.789	0.577	0.552
NB	0.735	0.75	0.614	0.632
CART	0.792	0.785	0.499	0.473
RF	0.897	0.791	0.804	0.568

ISSN: 2005-4297 IJCA

Table 5. Performance metrics for 'freshmen' dataset

Model	Acc		F1 Score	
	Training	Test	Training	Test
LR	0.818	0.81	0.722	0.707
NB	0.796	0.79	0.705	0.694
CART	0.802	0.789	0.707	0.682
RF	0.832	0.816	0.738	0.711

4.2 Identifying dataset characteristics using information gain measures

Table 6 and 7 compiles the information gain calculations using entropy and Gini index for each feature in 'tcc' and 'freshmen' dataset respectively. Each feature was assigned rank $R_{\{entropy,f\}}$ and $Rank_{(Gini,f)}$ in descending order of $IG_{\{entropy,f\}}$ and $IG_{\{gini,f\}}$ respectively. $R_{(dataset,f)}$ is the average of the two ranks $R_{\{entropy,f\}}$ and $Rank_{(Gini,f)}$ assigned to a feature.

Table 6. Information gain measures for 'tcc' dataset

Feature(f)	$IG_{\{entropy,f\}}$	$R_{\{entropy,f\}}$	$IG_{\{gini,f\}}$	$Rank_{(gini,f)}$	$Rank_{(dataset,f)}$
Contract	0.142	1	0.065	1	1
OnlineSecurity	0.093	2	0.047	3	2.5
Tenure	0.093	3	0.048	2	2.5
TechSupport	0.091	4	0.046	4	4
InternetService	0.08	5	0.04	5	5
OnlineBackup	0.067	6	0.033	7	6.5
PaymentMethod	0.064	7	0.036	6	6.5
DeviceProtection	0.063	8	0.031	8	8
MonthlyCharges	0.037	9	0.018	9	9

Table 7. Information gain measures for 'freshmen' dataset

Feature(f)	$IG_{\{entropy,f\}}$	$R_{\{entropy,f\}}$	$IG_{\{gini,f\}}$	$Rank_{(gini,f)}$	$Rank_{(dataset,f)}$
ScholarshipBracket	0.197	1	0.125	1	1
MarksCategory	0.116	2	0.067	2	2
HostelorTransport	0.1	3	0.062	3	3
FeePaidCategorized	0.091	4	0.052	4	4
LoanLetter	0.027	5	0.015	5	5
HomeTownType	0.022	6	0.014	6	6
PreviouslyStudied	0.005	7	0.003	7	7
QualifyingExam	0.002	8	0.001	8	8
AdmissionMonth	0.001	9	0.001	9	9

Top features giving maximum information gain for 'tcc' dataset were 'Contract', 'OnlineSecurity', 'tenure', 'TechSupport' and 'InternetService'. Top features for 'freshmen' dataset were 'ScholarshipBracket', 'MarksCategory', 'HostelorTransport',

ISSN: 2005-4297 IJCA

'FeePaidCategorized' and 'LoanLetter'. The ranks using entropy and gini index were observed same for both datasets, with 'OnlineSecurity' being the only exception.

4.3 Model behaviour in terms of feature importance

Owing to its best classification performance, RF model was selected for evaluating the proposed approach. Table 8 and 9 compiles the quantitative measure of importance given to a feature by the ML model and corresponding rank assigned for 'tcc' and 'freshmen' respectively.

Table 8. Feature Importance for tcc dataset

Feature(f)	f_{Imp}	$R_{\{Model,f\}}$
tenure	1.94	1
MonthlyCharges	1.689	2
Contract	1.634	3
OnlineSecurity	1.441	4
PaymentMethod	1.404	5
TechSupport	1.344	6
InternetService	1.327	7
OnlineBackup	1.317	8
DeviceProtection	1.257	9

Table 9. Feature Importance for freshmen students

Feature(f)	f_{Imp}	$R_{\{Model,f\}}$
ScholarshipBracket	1.694	1
FeePaidCategorized	1.198	2
HostelorTransport	1.14	3
AdmissionMonth	1.123	4
MarksCategory	1.12	5
HomeTownType	1.078	6
QualifyingExam	1.066	7
LoanLetter	1.061	8
PreviouslyStudied	1.032	9

Top features to which RF model outcome was sensitive included 'tenure', 'MonthlyCharges', 'Contract', 'OnlineSecurity' and 'PaymentMethod' for 'tcc' dataset. Similarly, for 'freshmen' dataset, the top features included 'ScholarshipBracket', 'FeePaidCategorized' ,'HostelorTransport', and 'MarksCategory'.

4.4 Measuring agreement between dataset characteristics and model behaviour

Table 10 and 11, gives a comparison of ranks assigned to features based on dataset measures and model behaviour.

ISSN: 2005-4297 IJCA

Table 10. Rank matrix for 'tcc' dataset

Feature(f)	$Rank_{(dataset,f)}$	$R_{\{Model,f\}}$
Contract	1	3
OnlineSecurity	2.5	4
Tenure	2.5	1
TechSupport	4	6
InternetService	5	7
OnlineBackup	6.5	8
PaymentMethod	6.5	5
DeviceProtection	8	9
MonthlyCharges	9	2

Table 11. Rank matrix for 'freshmen' dataset

Feature(f)	$Rank_{(dataset,f)}$	$R_{\{Model,f\}}$
ScholarshipBracket	1	1
MarksCategory	2	5
HostelorTransport	3	3
FeePaidCategorized	4	2
LoanLetter	5	8
HomeTownType	6	6
PreviouslyStudied	7	9
QualifyingExam	8	7
AdmissionMonth	9	4

Spearman's rank correlation was used to measure the degree of agreement between features considered important by our black box ML model and features rich in information gain as computed from the original dataset. Spearman's rank correlation coefficient value observed for 'tcc' and 'freshmen' dataset was r_{tcc} = .41 and $r_{freshmen}$ = .57 respectively. The positive values of the rank correlation coefficient indicate agreement between dataset measures and model behaviour. Comparing the correlation coefficients for 'tcc' and 'freshmen' datasets, it can be concluded that black box model in case of 'freshmen' dataset has incorporated characteristics of the dataset into its reasoning better than its counterpart for 'tcc' dataset. Referring to Table 3 and 4, a comparison of generalization of the RF model in case of 'tcc' and 'freshmen' dataset also validate our above conclusion.

4.5 Unboxing classification outcomes to facilitate trust of human users

To facilitate the trust of human users in ML-based solution, global surrogate models and local surrogate models were used. Global surrogate models help in explaining the global behaviour of an ML model. The local surrogate model helps in explaining the reasoning behind the classification outcome of a particular instance. The proposed framework allows on-demand explanations for prediction outcomes of the ML model. By evaluating these explanations for multiple instances, a human user is likely to trust the ML model if the reasoning of the model is in sync with the prevailing domain knowledge.

ISSN: 2005-4297 IJCA

Global surrogate model: Figure 2 and 3 shows a global surrogate model with depth fixed to two levels for 'tcc' and 'freshmen' dataset, respectively. For 'tcc' dataset, globally top significant features included 'OnlineSecurity', 'tenure' and 'InternetService'. Similarly, for 'freshmen' dataset, features selected for tree construction included 'ScholarshipBracket', 'MarksCategory' and 'FeePaidCategorized'.

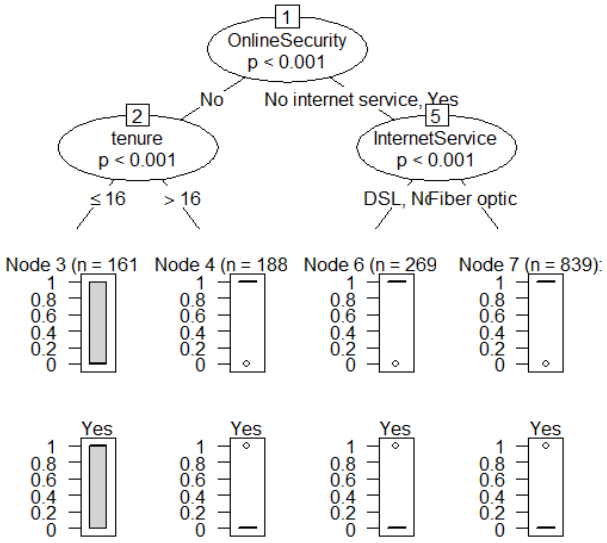

Fig 2. Global surrogate model for 'tcc' dataset

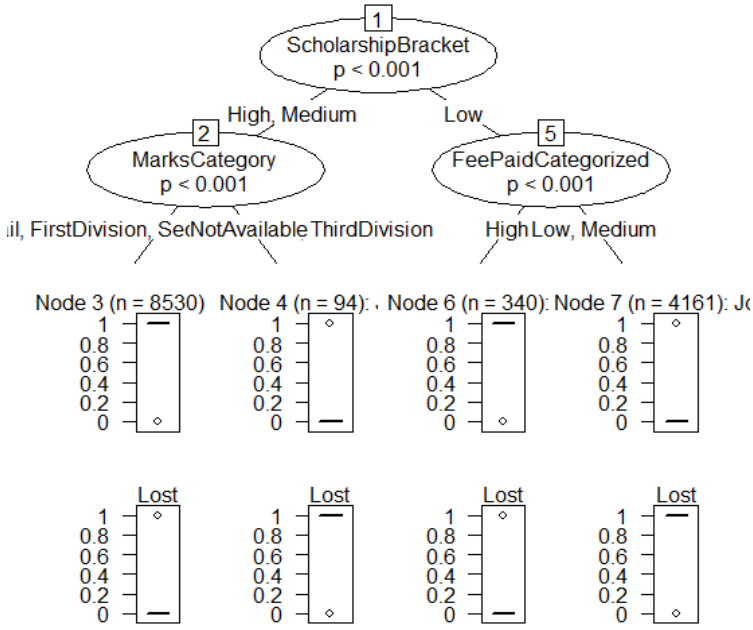

Figure 3. Global surrogate model for 'freshmen' dataset

Local surrogate models: Figure 4 and 5 shows local surrogate models using LIME. Blue coloured contributions (towards the right of the Y-axis) are supporting the outcome while red-coloured (towards the left of the Y-axis) contributions contradicts. The number of

ISSN: 2005-4297 IJCA

contributors was fixed to top 4 in each explanation. In case of 'tcc' dataset, the most significant contributors were 'Contract', 'tenure', 'InternetService' and 'TechSupport'. Similarly, for the instance from 'freshmen' dataset, 'ScholarshipBracket', 'MarksCategory', 'FeePaidCategorized' and 'HostelorTransport' were the top contributors.

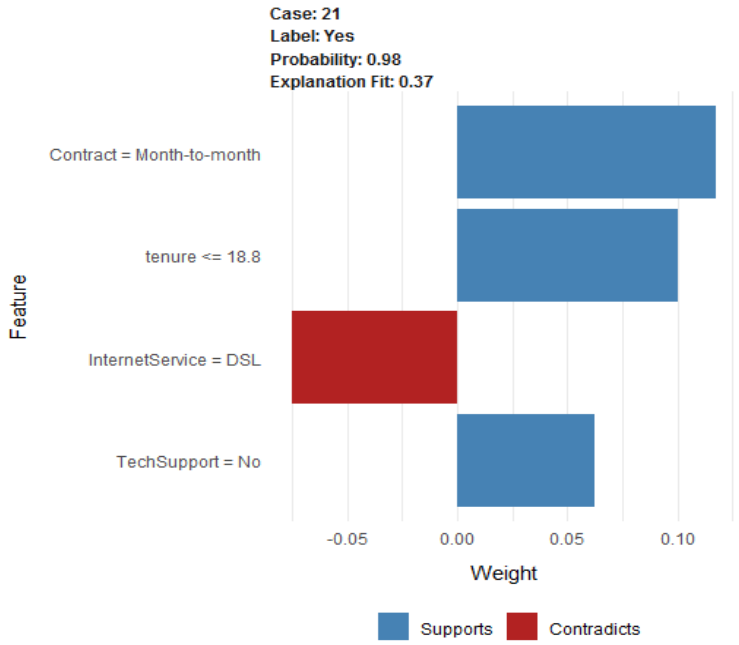

Figure 4. Unboxing classification outcome for 'tcc' instance

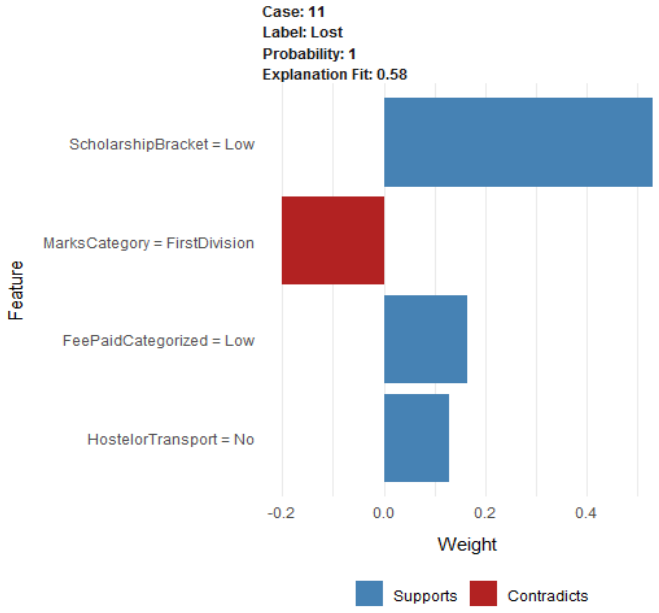

Figure 5. Unboxing classification outcome for 'freshmen' instance

For both 'tcc' and 'freshmen' datasets, most of the top features reported using surrogate models were also included among top features as per our actual RF model. These outcomes validate our understanding of model behaviour.

ISSN: 2005-4297 IJCA

5. Conclusion

The intuitive idea of taking agreement between important characteristics of the input dataset and model behaviour can be taken as a measure of the reliable learning process. The task of measuring this agreement quantitatively can be modelled as a 2-Judge n-participants rank correlation problem. Each feature used for learning is equivalent to a participant. The dataset and the learned ML model acts as two judges that assign a rank to each participant(feature) as per its importance. A positive value of the rank correlation coefficient indicates the agreement between the ML model and important dataset characteristics. The value of the rank correlation coefficient can be used to compare two ML solutions in terms of the degree to which an ML model has incorporated important characteristics of the input dataset into its reasoning.

Global and local surrogate models can help the human user understand the reasoning behind the behaviour of an ML model. Providing on-demand explanations of prediction outcomes help develop the trust of human users in the ML model. Even if, a user is not in agreement with the reasoning being used by the ML model, it may lead to the discovery of knowledge still unknown to the human experts. For Example, unknown interactions between predictors in problem domains like drug discovery. The proposed framework takes care of the reliability of the learning as well as interpretability needs of an ML-based solution.

There are multiple future research directions for this study. Chi-square and R-squared value can be explored as additional measures for dataset characteristics. Additional interpretability techniques can be used to understand model behaviour in a more comprehensive manner. This work can be extended to regression problems also.

6. Acknowledgement

We are thankful to Dr Ashok Sharma and Dr Santosh Kumar Henge, Department of Computer Applications, Lovely Professional University (LPU) for their valuable inputs through rigorous brainstorming. We also acknowledge the Kaggle team for providing access to Telco Customer-churn dataset.

7. References

[1] Tom M. Mitchell, "The Discipline of Machine Learning", 2006, CMU, http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf
[2] Houtao Deng," Interpreting Tree Ensembles with inTrees", arXiv:1408.5456v1 [cs.LG] 23 Aug 2014
[3] Satoshi Hara, Kohei Hayashi, "Making Tree Ensembles Interpretable", 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA.
[4] Hui Fen Tan, Giles J. Hooker, Martin T. Wells, "Tree Space Prototypes - Another Look at Making Tree Ensembles Interpretable", NIPS 2016 Interpretable ML for Complex Systems Workshop
[5] Gilles Vandewiele, Olivier Janssens, Femke Ongenae, Filip De Turck, Sofie Van Hoecke, "GENESIM: genetic extraction of a single, interpretable model", 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. arXiv:1611.05722v1
[6] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, Klaus-Robert Müllera," Explaining nonlinear classification decisions with deep Taylor Decomposition", Pattern Recognition 65 (2017) 211–222
[7] Wojciech Samek, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Klaus-Robert Müller, "Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation", Workshop on Interpretable Machine Learning for Complex Systems (NIPS 2016), Barcelona, Spain
[8] Erik ˇStrumbelj, Igor Kononenko, "An Efficient Explanation of Individual Classifications using Game Theory", Journal of Machine Learning Research 11 (2010) 1-18

ISSN: 2005-4297 IJCA

[9] David Baehrens, Timon Schroeter, Stefan Harmeling, "How to Explain Individual Classification Decisions", Journal of Machine Learning Research 11 (2010) 1803-1831
[10] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier", KDD 2016, San Francisco, CA, USA, http://dx.doi.org/10.1145/2939672.2939778
[11] Yotam Hechtlinger," Interpretation of Prediction Models Using the Input Gradient", arXiv:1611.07634v1 [stat.ML] 23 Nov 2016, NIPS
[12] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Model-Agnostic Interpretability of Machine Learning", ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA
[13] Sameer Singh, Marco Tulio Ribeiro, Carlos Guestrin," Programs as Black-Box Explanations", arXiv:1611.07579v1 [stat.ML], Neural Information Processing Systems (NIPS 2016), Barcelona, Spain
[14] L. Breiman, J. Friedman, C. J. Stone, and R.A. Olshen. (1084) "Classification and Regression Trees", CRC press
[15] L. Breiman (2001), "Random forests.", Machine learning, 45(1): 5–32
[16] https://cran.r-project.org/web/packages/entropy/entropy.pdf
[17] https://cran.r-project.org/web/packages/reldist/reldist.pdf
[18] Zachary C. Lipton. (2016) "The Mythos of Model Interpretability", ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA
[19] Kiri L. Wagstaff. (2012) "Machine Learning that Matters" Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012
[20] https://christophm.github.io/interpretable-ml-book/

ISSN: 2005-4297 IJCA