MODEL INTERPRETABILITY VIA INTERACTION FEATURE DETECTION USING ROUGHSET IN A GENERALIZED LINEAR MODEL FOR WEATHER PREDICTION IN KENYA.

Isaac Kega[1], Lawrence Nderu[1], Ronald Mwangi[1], and Dennis Njagi[1]

[1]Affiliation not available

May 25, 2023

Abstract

Feature interaction is a crucial concept when it comes to machine learning interpretability. These interactions make it easier for the person using a machine learning tool to know how the model made its prediction. Intrinsic models such as the Generalized linear models pull their interpretability aspects from detecting feature interactions on the data. However, these models do not search the whole sample space for interactions and assume all interactions to the target feature are identical. A hybrid model combining Rough Set and Generalized linear models is proposed. Rough Set uses the concept of information granulation and approximations of regions with meaningful information to find feature interactions in the whole sample space. The detected features will then be modeled onto a Generalized linear model for prediction purposes. The dataset used in the experiment was drawn from wunderground.com online weather site, specifically the Kariki Farm online weather station. The proposed methodology for the research is the CRISP-DM method.

Keywords: Machine interpretability, Rough Set, Generalized linear model, weather prediction.

INTRODUCTION

Interpretable machine learning methods can be used to discover new knowledge, debug and justify the models and their predictions, control the model, and improve it. (Molnar, C., Casalicchio, G., & Bischl, B, 2020). While a lot of machine learning research has been focusing on the predictive performance of the models. One lagging aspect, though not by much of it being left utterly behind, is the interpretability of the models used in making the prediction. (Molnar, C., Casalicchio, G., & Bischl, B, 2020). The diagram below shows some machine learning models' interpretability vs. accuracy tradeoffs. It offers a need to improve on interpretability aspects of machine learning. From the chart, we can see a lot of the black box models are models that consider nonlinear relationships in their structure. In contrast, the interpretable ones are linear in design and have well-defined relationships. As you can see, all linear regression models are highly interpretable while Deep learning models aren't; hence need to bridge this gap.

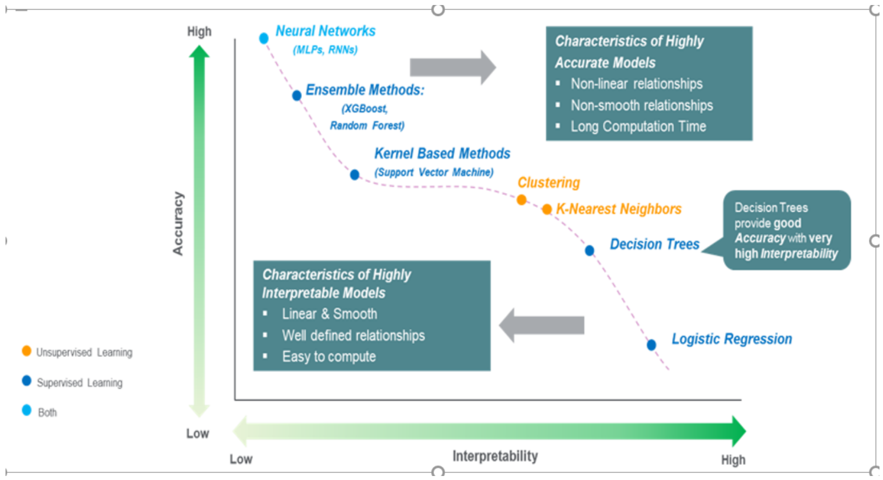

Figure 1: Interpretability vs. accuracy tradeoff in Machine Learning

LITERATURE REVIEW

There is a great need to understand how things occur. Hence the need for comprehensibility or interpretability. Interpretation can be said to be the action of explaining something to know how that aspect works, why the part behaves in a particular manner and any impact that the element in question can have on attributes depending on it. In machine learning, interpretability is highly important nowadays as these models run vital critical processes worldwide in financial, academic, and political scenarios.

Why is interpretability important? From the literature reviewed, interpretability is essential based on TWO reasons. One aspect of being a human being is that we need to put an explanation to every event that occurs in our lives. This need thus has driven the urge to be able to explain certain causal circumstances and why they informed a particular decision. This need is equally essential in machine learning as we need to explain why a particular model chooses a specific prediction instead of another. Thus the interpretability of machine learning models can aid human beings in learning new aspects and also find meaning in their decisions. (C Molnar 2019)

The other is handling uncertainty and biases in decisions. Much research has developed new models to handle uncertainty and bias in machine learning models. Certainly, interpretability is an important aspect that can be used to address discrimination and uncertainty. Machine learning models pick preferences from training data caused by missing values in some datasets, which leads to uncertainty in these models when it comes to predicting the outcomes based on the training data they have. This makes the models unstable and gives outrageous predictions that cannot be explained. Thus interpretability can be used to handle this issue and offer a reprieve for models and their developers. (C Molnar 2019)

Review of literature by Christoph Molnar in his book "Interpretable machine learning," we see that interpretable machine learning models can be classified according to FOUR fronts. The figure below illustrates the classification of interpretable machine learning models.

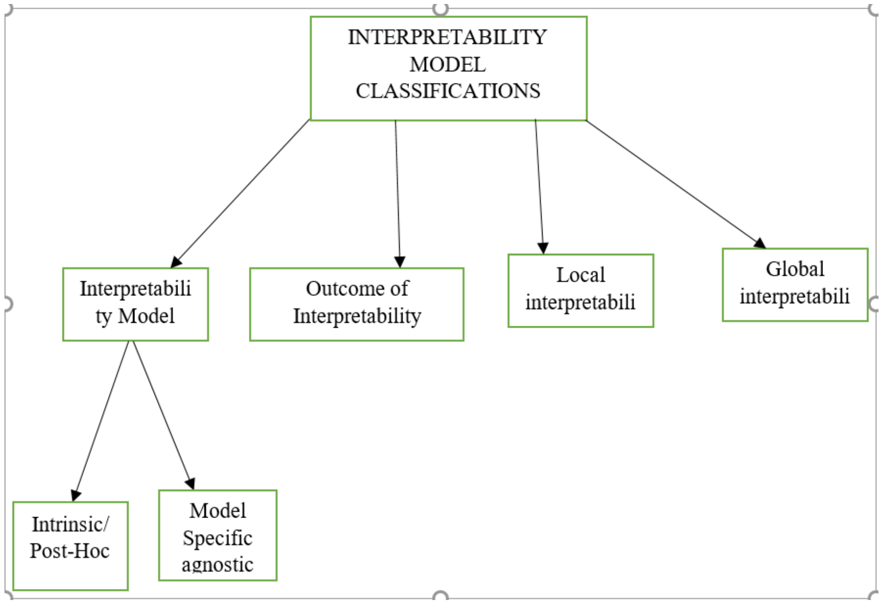

Figure 2: Interpretability of model classification

Interpretability model: this classification is based on two methods which are intrinsic/post-hoc methods and model-specific agnostic models. Intrinsic is applied intrinsically interpretable models; these are models whose structure is simple and thus can easily be interpreted, while post-hoc models are used to analyze a specific model after training. A good example is the permutation feature. Model-specific/model agnostic method is another sub-domain under the interpretability model method. They are model-specific and limited to a specific machine learning model, while model-agnostic methods can be applied to any machine learning model after training.

Local and Global interpretability models depend on the scope of the trained model, which includes aspects like knowledge of the model, the algorithm used, and the data that shall be used in the prediction. This is in the case of global, while local interpretability chooses a particular instance during the execution of the model when a prediction was made. Thus the model is interpreted in that instance. Global explanations give a holistic view, while local descriptions examine single instances in the model. (C Molnar 2019).

GENERALIZED LINEAR MODEL

Generalized linear models are an offshoot of the interpretable model linear regression. It predicts a target as a weighted sum of inputs. The way linear regression makes interpretability easy is by viewing the learned relationships' linearity. This is achieved through modeling the dependence of the target variable y on some features of X. It's mainly used to tackle quantitative problems by statisticians and computer scientists (C Molnar, 2019).

A GLM is one of the extensions of the linear model to model nonlinear outcomes, the essential defining feature of a GLM is to allow non-Gaussian outcome distributions and connect the distributions and the weighted sum through a nonlinear function. Thus a GLM can be modeled to give a categorical outcome and a count outcome, which a linear regression model cannot produce (C Molnar, 2019).

GLM consists of **THREE** essential parts, namely:

Random Component – refers to the probability distribution of the response variable (Y), e.g., normal distribution for Y in the linear regression or binomial distribution for Y in the binary logistic regression. Also called a noise model or error model.

Systematic component - specifies the explanatory variables (X_1, $X_2 \ldots X_k$) in the model, more specifically their linear combination in creating the so-called *linear predictor* ; e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ as we have seen in linear regression.

Link Function, η *oρ* γ (μ) - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Y_i)) = E(Y_i)$ for linear regression or $\eta = logit(\pi)$ for logistic regression.

ROUGHSET AS A FEATURE INTERACTION DETECTION MODEL

Rough Set theory is a knowledge discovery method highly applied in relational databases. Rough Set is a machine learning model that bases its functionality on the information granulation of the data it is working on. That is, it seeks to identify the interactions in the data even if it has incomplete or no prior information. Professor Pawlak first introduced it in 1982. Rough sets can be divided into two parts; the first part forms the concepts and rules through classification, while the second concerns knowledge discovery through target classification. Rough sets have been used in several types of research when coupled with machine learning methods; they have been used in preprocessing problems, feature selection, and instance selection (Bello, R 2017). Rough Set theory's fundamental concepts are as explained below:

Indiscernibility Relation : is the relation between an object in a rough set where all the values are identical to the subset of the considered attributes (Rissino, S. et al. (2009))

Let A, P [?] A, the indiscernibility relation IND (P), can be defined as IND (P) = {(x, y) [?]U ×U: for all a[?] P, a(x) = a(y)}

A set is a grouping of objects which contain similar characteristics. (Rissino, S. et al. (2009))

When the boundary region is a non-empty set that is *B(X)* [?]*B(X),* then the set is Called a Rough Set.

Approximations : are based on the THREE regions of the rough set theory, mainly lower approximation, upper approximation, and boundary approximation. (Rissino, S., et a (2009))

A lower approximation of a subset can be defined as the set of objects that positively belong to the target set. Let B [?] C and X [?] U, the B-lower approximation set of X, be the set of all elements of U, which can be with certainty classified as elements of X.

B(X) ={*x*[?]*U: B*(*x*) [?] *X*}

Upper Approximation can be defined as a set of objects which possibly belonging the target set.

B(X) = {*x*[?]*U: B*(*x*) [?] *X*[?] φ }

Boundary Approximations can be classified as the collection of elementary sets of objects that cannot be decisively classified into *X* in *B.*

$BN_B(X)$ = *B(X)* -*B(X*

Decision Table / Information system: this is the primary mode of storing data in rough sets and represents input data gathered from the domain or environment in which the rough sets will be implemented. (Rissino, S. et al. (2009))

Reduct: This process in Rough set theory involves dimensionality reduction through removing redundant or irrelevant attributes. This process is also the result of the feature selection process of the roughest technology, and the end products are decision tables called Decision Reducts. The Quick reduct method and greedy heuristic method can be used for reduct generation, which combines the greedy search method with a heuristic function such as entropy to find a minimal subset of features necessary for decision-making.

METHODOLOGY

The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is proposed for use in this research work. The method is preferred for most data mining projects being done today. The choice in methodology is due to the aspect of the CRISP-DM being independent of any industry domain and technology that can be used to implement the data mining solution. It comprises five steps to follow while doing the work.

In the first step, the researcher has to set goals for the project or work being done; this will help develop the objectives of the work being done. Second step, the researcher has to specify the data and sources they will use in their work. The second step is essential as it defines the type of data that shall be used in the project, whether it meets the requirements of the project, and achieves the project's intended results. The third step would be data preparation and transformation to make the data ready and usable for the project. The fourth step would be to model the algorithm for my proposed solution and use it on the data I had gathered in the previous three steps. The fifth step will be to evaluate the results of my proposed model based on the accuracy and interpretability of the model I had deduced in step Four. Lastly, the sixth step would be to have a conclusive end to the research based on the evaluation findings of step 5 and comparison with other methods in the same area (Wirth et al., 2000).

EXPERIMENT SETUP

In this experiment, a Rough set shall be used to determine the interaction terms of the variables in the data. The data was from wunderground.com, specifically the Kariki_Farm weather station in Juja. The data keeps on getting updated because it is real-time data. The dimension for the information is as follows:

S/No	Dataset	Dataset Source	Dataset Description Dimension	Dataset Description Instances
1	Kariki Farm PWS	wunderground.com	16	1620

Table 1: Kariki Farm Dataset Dimension from wunderground.com

First, the data was loaded and preprocessed to ensure it was clean and had no missing values. Next, we used Rough Set to detect the interaction terms in the dataset. The process was achieved by first discretizing the data, which changes numerical representations to nominal, which was necessary to enhance the evaluation and management of data. Discretization uses a data transformation procedure involving finding and cutting data sets and dividing the data into intervals. Values lying within an interval are then mapped to the same value. Doing this process will reduce the size of the attribute value set (Hassanien, A. E., Abdelhafez, M. E., & Own, H. S. (2008)).

Next, the indiscernibility relation was used to determine which variables in the dataset are indiscernible from the rest. From this relation, we can now deduce the lower, upper, and boundary approximations, which determine the lower, upper, and boundary regions, respectively. The lower region represents attributes/variables belonging to the subset of interest. After deducing the approximations, the next step is to formulate the reduct (feature subset) from the lower/positive region of the approximations; the method employed here will be the greedy heuristic method for feature selection which is a wrapper feature selection algorithm. (Janusz, A., Ślezak, D. (2014)).

The experiment framework is shown in the diagram below:

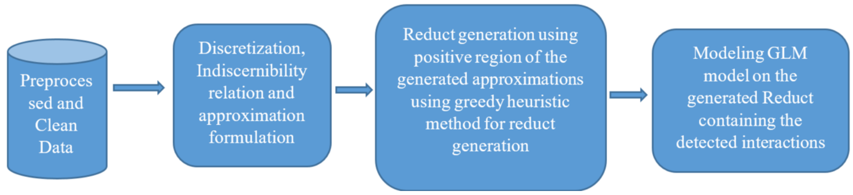

Figure 3: Proposed experiment framework using Rough Set theory to detect interaction terms for a Generalized Linear Model.

FINDINGS AND DISCUSSION

The experiment was carried out in two phases:

In this first phase, the data was modeled using the GLM model based on the Binomial random component, which specified the use of the Binomial probability distribution function for the response variable. The link function used here was the Logit function, which is part of the logistic regression to model categorical variables because, in the experiment, we are to determine whether, based on the available variables, it will rain or not. The training accuracy for the model was 84%, while the testing accuracy was 83%. The experiment results are as follows

Accuracy	Sensitivity	Specificity	Kappa
0.8294	0.8889	0.700	0.593

Table 2: Experiment result **GLM model fit on Original Kariki_Farm data without interaction detection using RST**

Confusion matrix for the above values after the original experiment

Predicted/ Actual	No	Yes
No	232	36
Yes	29	84

Table 3: Confusion matrix **GLM model fit on Original Kariki_Farm data without interaction detection using RST**

The critical attributes necessary for prediction from the data provided as determined by the fit logistic regression model were Low_temp, Dewpoint_high, Windspeed_high, and windspeed_avg.

In the second experiment, we used Rough Sets to detect interaction terms for the model. Here the data was loaded, and Rough Set theory was used to determine the critical features using lower and upper approximations. After deducing the approximations, the next set of operations was to formulate the reduct (feature subset) from the lower/positive region of the approximations; the method employed here was the greedy heuristic method for feature selection which is a wrapper feature selection algorithm. The reduct formulated had 12 important variables: High temp, Avg temp, low temp, high dew-point, low dew-point, high humidity, low humidity, high wind-speed, avg wind-speed, high pressure, and low pressure. The accuracy was 84.25 %. This shows an increase in testing accuracy by 1.31%. The important attributes necessary for prediction from the data provided as determined by the fit logistic regression model were Dew-point_high, Winds-peed_-high, wind-speed_avg, humidity_high, and High. hpa. As in this case, humidity and high pressure were not considered essential attributes in the prediction model in the first experiment.

Accuracy	Sensitivity	Specificity	Kappa
0.8425	0.8889	0.8000	0.6812

Table 4 Experiment results **GLM model fit on Kariki_Farm data with interaction detection using RST.**

Confusion matrix for the above values after the original experiment

Predicted/ Actual	No	Yes
No	230	29
Yes	31	91

Table 5: Confusion matrix **GLM model fit on Kariki_Farm data with interaction detection using RST**

From the experiment, we see that the GLM model, modeled using the detected interaction terms from the Rough Set theory method, performed better in accuracy compared to the model without the detected interactions. Seeing these interactions helped determine the essential features necessary for predicting the weather, reduced the dimensionality of the data, and ultimately improved the accuracy of the GLM model. The table below shows these results:

Model Specification	**Accuracy**	**Sensiti**
GLM model fit on Original Kariki_Farm data without interaction detection using RST	0.8294	0.8889
GLM model fit on Kariki_Farm data with interaction detection using RST	0.8425	0.8889

Table 6: Experimental differences between the TWO experiments done

Table 6 metrics can be summarized in the chart below, depicting the performance of the GLM model fit with interactions detected via the Rough Set Theory model vs. a classical GLM model without interaction detection. We see the proposed model had better accuracy, sensitivity, specificity and Kappa measures than the GLM model without interaction detection.

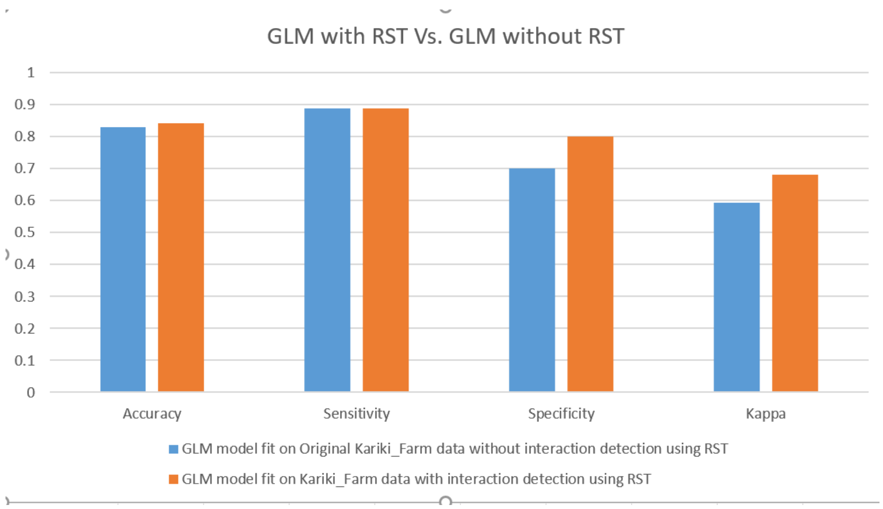

Figure 4: GLM with interaction detection via Rough Set Theory vs Classical GLM without interaction detection

DISCUSSION

Rough Set Theory is a mathematical framework that deals with incomplete or uncertain data by analyzing a given dataset's lower and upper approximations. In this work, Rough set theory improved prediction and interpretability of the Generalized linear model through feature selection which ultimately led to detecting key features that influenced the target variable, these detected features were fed to the Generalized Linear model which improved the accuracy and also improved interpretability of the model by removing non important features.

CONCLUSION AND RECOMMENDATIONS

In conclusion, the experiment showed that using the Rough Set theory for interaction detection and feature selection improves the logistic regression model's accuracy and reduces the dataset's dimensionality, thus reducing the time required to execute the logistic model. The next part of the experiment will be to model an association rule mining method on the generated reduce and produce decision rules using the technique.

REFERENCES

1. Bello, R., & Falcon, R. (2017). Rough sets in machine learning: A review. In *Thriving Rough Sets* (pp. 87-118). Springer, Cham.

2. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

3. Hassanien, A. E., Abdelhafez, M. E., & Own, H. S. (2008). Rough Sets Data Analysis in Knowledge Discovery: A Case of Kuwaiti Diabetic Children Patients. *Advances in Fuzzy Systems* .

4. Janusz, A., & Ślezak, D. (2014). Rough set methods for attribute clustering and selection. *Applied Artificial Intelligence* , *28* (3), 220-242.

5. Molnar, C. (2019). *Interpretable machine learning* . Lulu. com.

6. Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning–A Brief History, State-of-the-Art, and Challenges. *arXiv preprint arXiv:2010.09337* .

7. Richardson, A., & Rosenfeld, A. (2018, July). A survey of interpretability and explainability in human-agent systems. In *XAI Workshop on Explainable Artificial Intelligence* (pp. 137-143).

8. Rissino, S., & Lambert-Torres, G. (2009). Rough set theory—fundamental concepts, principals, data extraction, and applications. In *Data mining and knowledge discovery in real-life applications* . IntechOpen.

9. Shi, L., Weng, M., Ma, X., & Xi, L. (2010). Rough set-based decision tree ensemble algorithm for text classification. *Journal of Computational Information Systems* , *6* (1), 89-95.

10. Vijaya, J., & Sivasankar, E. (2018). Computing efficient features using rough set theory combined with ensemble classification techniques improves customer churn prediction in the telecommunication sector. *Computing* , *100* (8), 839-860.

11. Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). London, UK: Springer-Verlag.

12. Widz, S., & Ślezak, D. (2012). Rough set-based decision support—models easy to interpret. *Rough Sets: Selected Methods and Applications in Management and Engineering* (pp. 95-112). Springer, London.