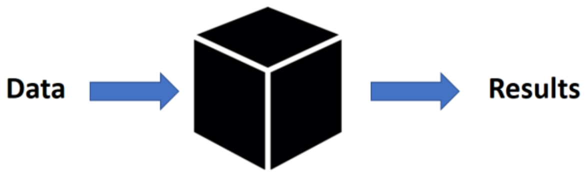
Validating the Black Box – Model Interpreters

Eoin Ó Baoighill, FSAI
Eamonn Phelan, FSAI, FIA, CERA

Models are becoming more sophisticated as insurers strive to maximise financial performance. As sophistication grows, interpretation of models becomes more difficult.



Insurers use models to predict the future, whether it is for future reported claims, ultimate claims costs or policy sales. Models at varying levels of sophistication are used. Businesses are sometimes reluctant to move to more sophisticated models as models can be difficult to interpret and trust.

Predictions from models are not always intuitive. Validation exercises can be undertaken to prove that they are reliable, but additional work is needed to ensure that they are doing what was expected of them.

There have been some public examples where models have not performed as expected. Gender bias or racial bias, for example, can occur unintentionally where input data is not representative. Changes in data collection over time can also invalidate results.

Unlike humans, a well-coded and well-explained algorithm may be able to clearly display its own biases. We need to understand our models better and to interpret what they are doing.

What is Model Interpretation?

Model interpretation is understanding how your models work and why they are accurate. If a model is interpretable, the predictions from that model should be intuitive to the modeller.

Model interpretation can be done at a global level, where the global behaviour of a model is explained. It can also be done at an individual prediction level for local explanation, where the factors driving the predictions for a single observation of interest can be isolated.

Why Interpret Models?

Model validation and interpretation are key and distinct steps in the data science workflow and are essential before deployment. Many tests can be done to prove that a model is predictive, without the need to understand exactly what the model is doing.

Key Steps in model interpretation

GDPR/Compliance	Trust
Model Improvements	Changing Behaviour

Models must be trusted before they can be deployed. Interpreting how model predictions are made gives reassurance that the model is working as expected.

In order to improve models, we need to understand how the existing models work. Once we understand how the model works and which data fields are most important, additional feature engineering and data collection can be influenced by the results of model interpreters.

Models are often used to predict behaviour. However, they can also be used to identify ways of changing behaviour. Telematics is a good example of this, where drivers' habits can be identified and remediated based on the results of a model. However, care needs to be taken to ensure that correlation is not interpreted as causation within the model.

The GDPR recitals, where the regulation is explained, reference a Right to Explanation of decisions made. While the recitals are not written into law, companies may need to explain decisions taken by their models.

Insurers must ensure that the models do not adversely introduce discrimination into pricing models. For example, in EU markets, gender cannot be used to price anymore.

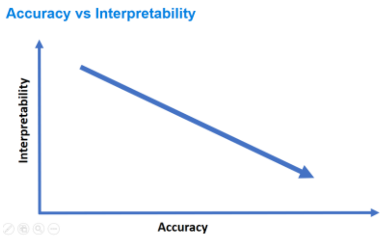
Requirements from Model Interpreters

There are several desirable qualities that model interpreters should have:

I. The key factors used in the model should be clear. The relationship between the input variables and the model predictions should also be easily understood.

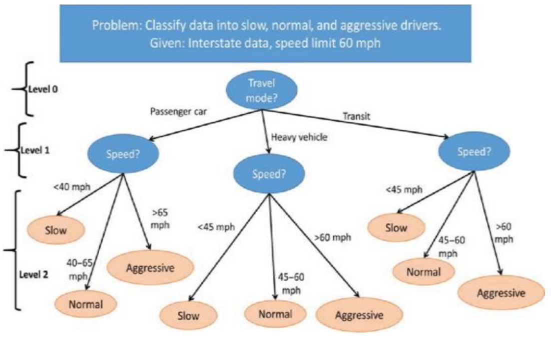
II. Both the global model behaviour and the model predictions at an individual observation level should be understood. Some model interpreters work at a local level. This means that they are representative for a given model point, but do not necessarily reflect the complete model. Regional explanations, where predictions for groups of observations can be explained, can provide a balance between local and global explanations.

Options for Interpreting Models



Simpler, and more interpretable, models often have lower predictive power. While these may be suitable for some tasks, insurers looking to maximise financial return need to consider more complex models for some applications.

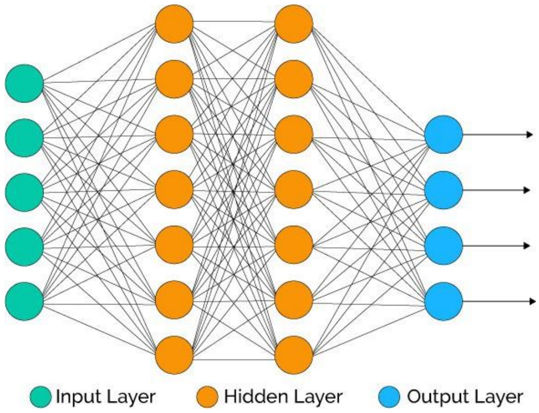
Decision trees are one of the more transparent algorithms with each node denoting a test on an attribute of the dataset, each branch representing an outcome of the test, and each leaf node (terminal node) holding the classification label or the regression outcome.



Source: Bhavsar, Parth & Safro, Ilya & Bouaynaya, Nidhal & Polikar, Robi & Dera, Dimah. (2017). Machine Learning in Transportation Data Analytics

In contrast, neural networks lend themselves to a more complex environment, modelled loosely on the human brain. They interpret data using various labelling and clustering techniques. The patterns recognised are numerical and contained in vectors. These vectors are then used to translate data such as images, sound or text. Neural networks with more than three layers are known as deep neural networks.

Neural networks consist of three layers and are roughly structured as follows:



Source: https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Several techniques are available to assist in interpreting the models, regardless of the complexity of the modelling techniques used.

Interpretable Models

Many solutions can be achieved using more straightforward models that are simpler to interpret.

Decision trees may have limited predictive power, but are easily explainable. They can assist in the segmentation of customers into easily defined groups. Feature importance can easily be judged by reviewing the splits within the decision tree.

Generalised regression models have been used successfully for many years. Coefficient values show the impact of each factor on the predicted values. These models are still widely used in predicting claims cost models.

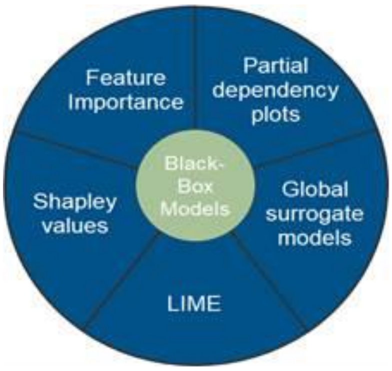
Simpler models are often more appropriate where there is limited data available, and where expert judgment is key.

Black Box Models

Generalised Linear Models have been widely used in insurance for many years. While they give good interpretability and predictive power, there has been a drive to improve on model accuracy. More complex models, such as random forests and gradient boosting methods are becoming more common as insurers seek to maximise predictive power.

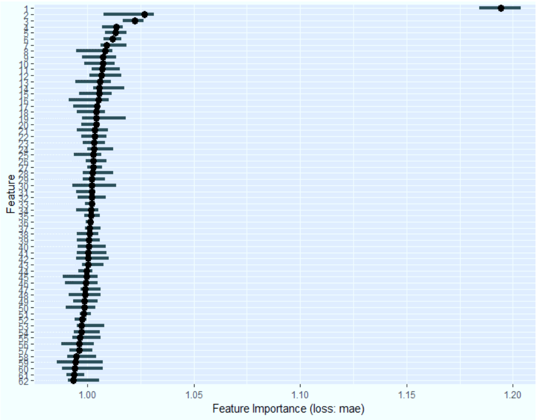
Model Validation Methods

There are a number of available validation methods for black-box models. We have explored the main ones in this section.



Feature Importance Plots

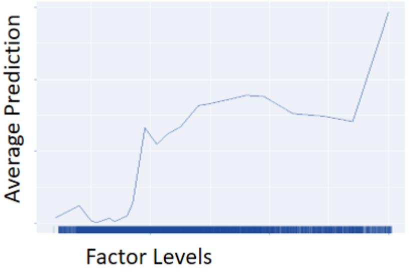
Feature importance plots map out the most important factors in the data.



There are several ways to do this. The most common is to shuffle the field values within a factor and to calculate the increase in the model's prediction error. If shuffling its values does not change the model error, the factor must not be important. Conversely, if shuffling the values increases the model error, the factor must be important because the model relied on it for the prediction.

Feature importance is easy to read and communicate. It does not comment on the direction of the relationships, or whether they are linear or non-linear.

Partial Dependency Plots



Partial dependency plots show the marginal effect that one or more features have on the predicted outcome. Model predictions are evaluated while varying the selected features.

The results are intuitive and easy to explain. However, there is an assumption that factors are not correlated. Sometimes looking at one factor at a time can be misleading. Accumulated Local Effect (ALE) plots can be used instead to mitigate the issues with correlations between factors. ALEs show how model predictions change as the value of a feature changes in a narrow range around the current value of that feature.

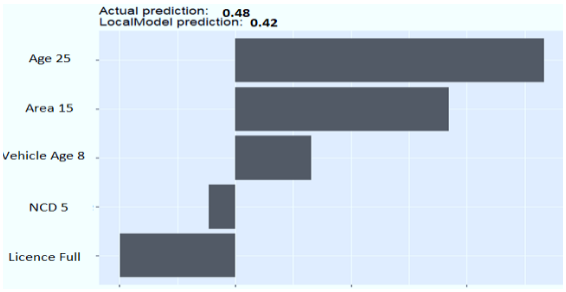
Partial Dependency Plots do not take heterogeneous effects into account, because of the average achieved. One solution to this is to use individual conditional expectation (ICE) plots. These show how the predictions of individual observations change as the value of a factor changes.

Global Surrogate Models

Global surrogate models are trained using the predictions of a black box model. The surrogate model is an interpretable model that can then explain elements of the black box model. Sparse linear regression, sparse logistic regression and decision trees that are not deep are often used as their results are easily interpretable.

Local Interpretable Model-Agnostic Explanations (LIME)

Local models are similar to Global Surrogate models, but focused on individual predictions instead of the global model. LIME is an example of one of these. Data features and predictions from the model are modelled using a simple model, typically a linear regression model. LIME attempts to explain a single observation at a time. Rather than using all available data in the same way that the black box model did, sample data is weighted according to their proximity to the individual data point. K-LASSO (a type of linear regression) is used, and the user selects the number of features to include in the regression model. The outputs show the most important features that contribute to the prediction for that data point.

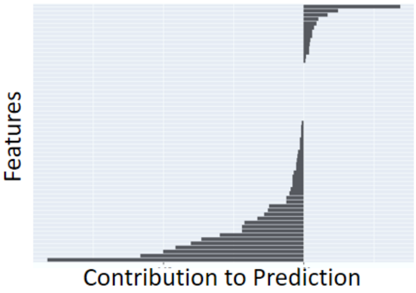


A number of simplifications are made within LIME. The explanation shows the key factors, but does not show the contribution from all factors. There is an assumption that a linear model or a decision tree is appropriate. In addition, data is weighted according to the distance between the observation of interest and other data available. These assumptions may not also be appropriate. The randomness in the sampling procedure introduces more uncertainty in the interpretations. The quality of interpretations can vary across different input data points.

Shapley Values

The Shapley value is the only explanation method with a solid mathematical theory to support it, The Shapley value is a concept in game theory introduced in 1953 to assign a fair distribution to a group of players that work together. This concept has been applied to data science to allocate the contribution from each factor in a model to the final predicted value.

The average predicted value across all data is compared with that for a single observation. The contribution from each factor in the model on the difference is calculated. This process can be run across all observations and aggregated to get a complete distribution of the prediction by feature.



LIME typically reviews a smaller number of factors, so Shapley should produce a more comprehensive and accurate analysis. It does come at a processing time cost, with Shapley analyses taking longer than LIME to run. In practice, sampling is used to reduce computing time, although that sampling introduces another source of uncertainty.

The best solution depends on the task. Sparse explanations (such as those that LIME produces) can sometimes be preferred, while other use cases require that a more comprehensive explanation is given.

Conclusion

Model interpretation is an important part of the data science workflow. There are many tools that can help. A typical analysis will involve the deployment of most of these tools, as specific strengths and weaknesses are associated with each one.

How Milliman can help

At Milliman, we have been actively working with our clients for many years to effectively harness the power of data science in order to help meet their business needs.

We can assist you with all aspects of your data science initiatives including providing advice on:

- Best practice frameworks for data science processes
- Model interpretation
- Model validation
- Collection and processing of data
- Identifying applications for data science techniques
- Identifying suitable tools and techniques for particular circumstances
- Implementing solutions
- Understanding the implications of results
- Constraints and practical challenges

For further information please contact any of the consultants below or your usual Milliman consultant.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT
Eoin Ó Baoighill
eoin.obaoighill@milliman.com

Eamonn Phelan
eamonn.phelan@milliman.com

The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.