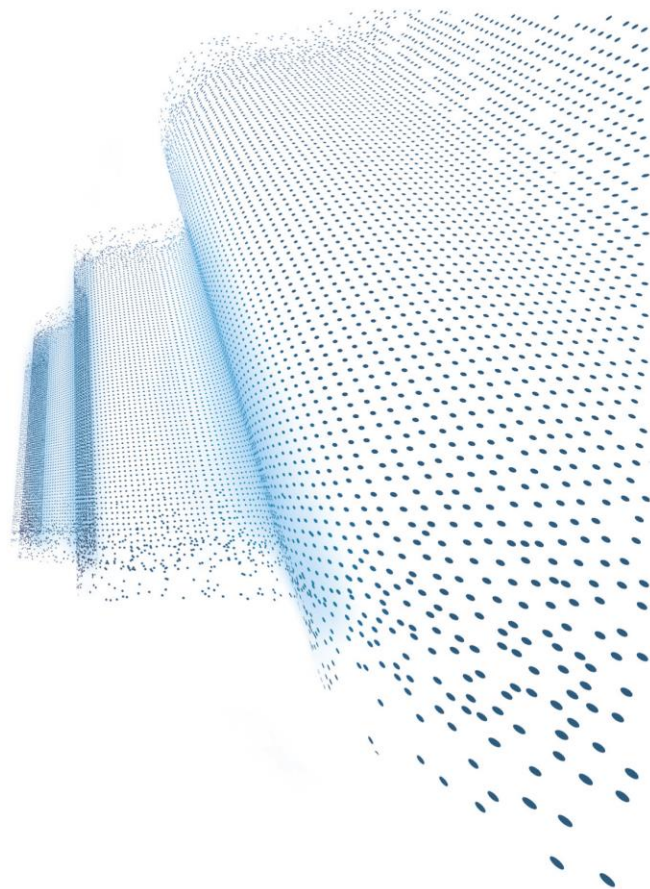




Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney





Model interpretation & explanation

Josh Jaroudy – Taylor Fry



Answering *how* and *why*

How does the model form its predictions?

Why did the model make a particular prediction?



Who cares?

Internal stakeholders

- Modellers & reviewers / audit
- End-users and others affected by the model (e.g. the sales/marketing team for a pricing model)

External stakeholders

- Regulators
- Customers
- Competitors

Models require trust

Do you **trust** the model to deliver the **right business outcomes** if deployed?

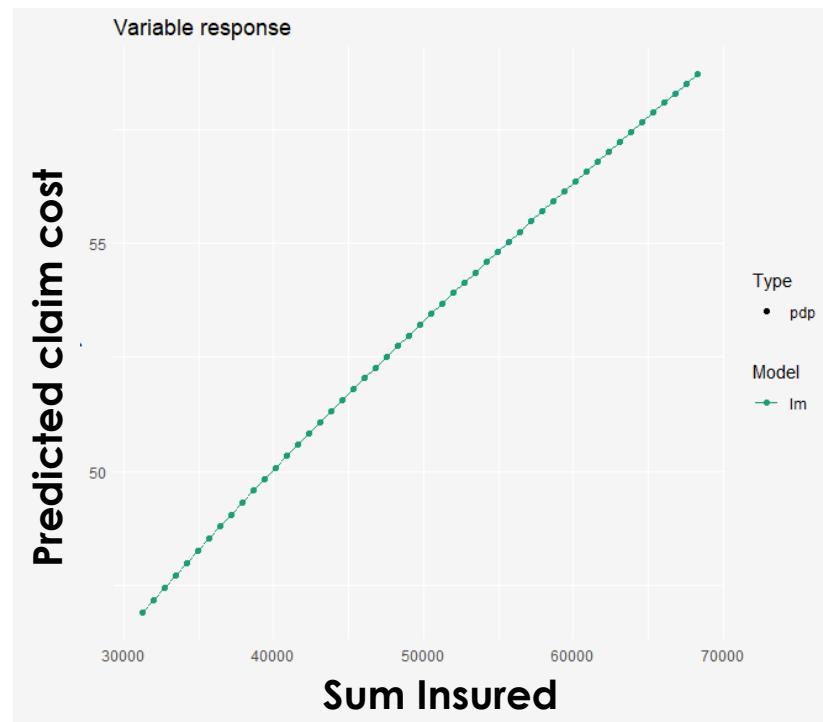
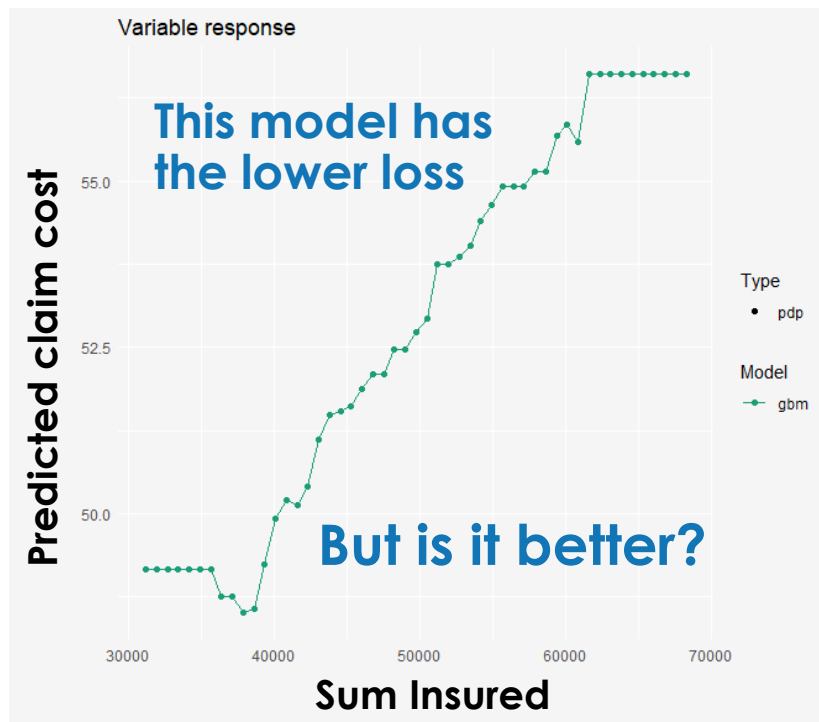
- Are the key modelled relationships likely to **persist into the future**?
- Are the outcomes **fair and reasonable**?
- Are there **reputational risks**?
- Does the model **comply with regulations** or support such compliance (e.g. 'Right to an explanation')



Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



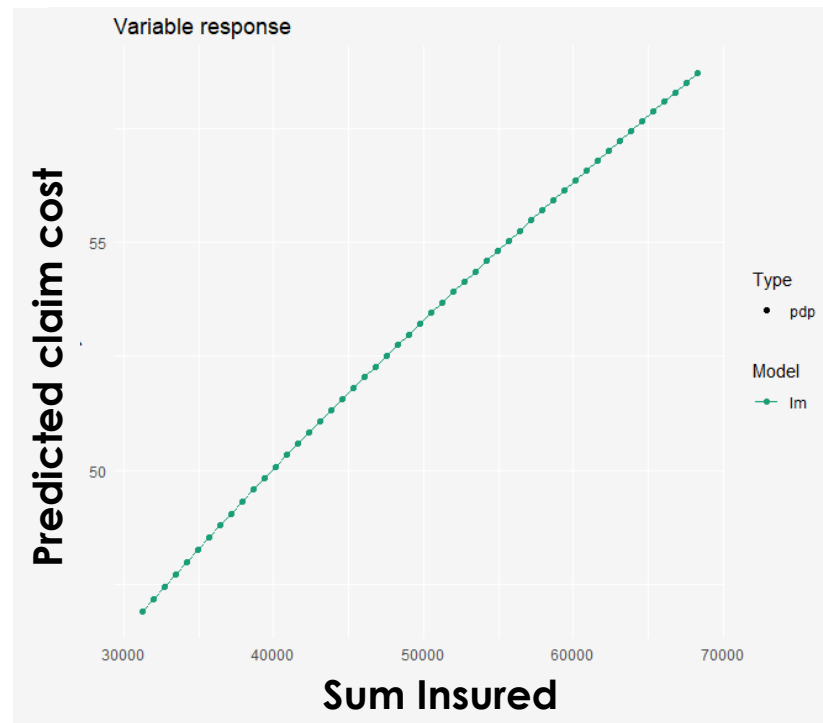
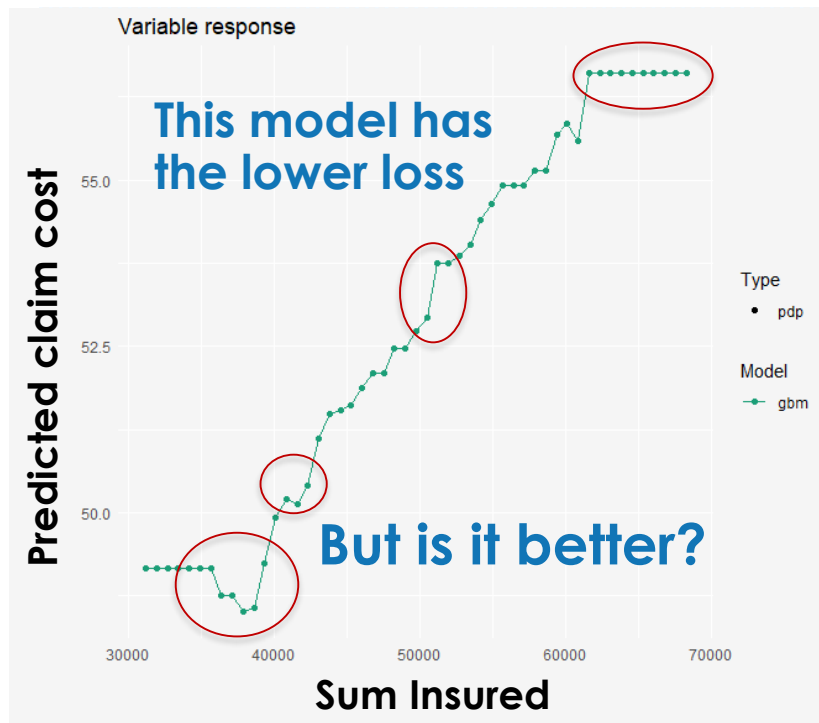
Models fit to simulated data, loss calculated out-of-sample



Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



Accuracy isn't everything



Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



Actuaries
Institute

Understanding the model structure



Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



DALEX



Descriptive m**A**chine
Learning **EX**planations

A package to **explain black-box model** predictions

Implements a variety of
methods for **global and local** explanations

Available on **CRAN** for **R**

github.com/pbiecek/DALEX

```
# Define the explainer
```

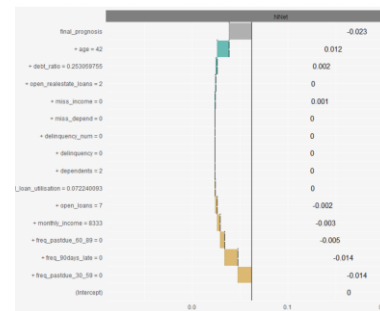
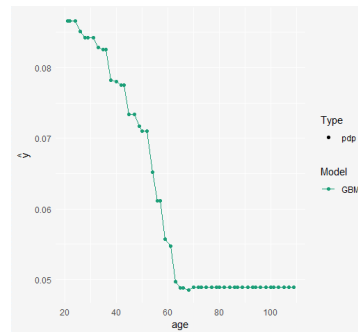
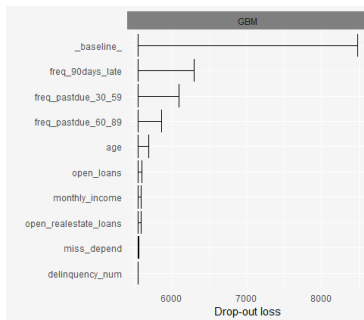
```
expl <- explain(model, data, y, predict_fun, link)
```

```
# Then produce the explanations
```

```
variable_importance(expl, loss_fun)
```

```
variable_response(expl, variable)
```

```
prediction_breakdown(expl, obs)
```





Example: credit default

- Data from Kaggle: Give Me Some Credit
<https://www.kaggle.com/c/GiveMeSomeCredit/data>
 - **Target:** did the individual experience financial distress in next 2 years
 - **Predictors:** age, income, number of dependents, number of times 30-59/60-89/90+ days past due in last 2 years, unsecured loan utilisation, number of open loans
- Models fit: a GBM and a deep neural network



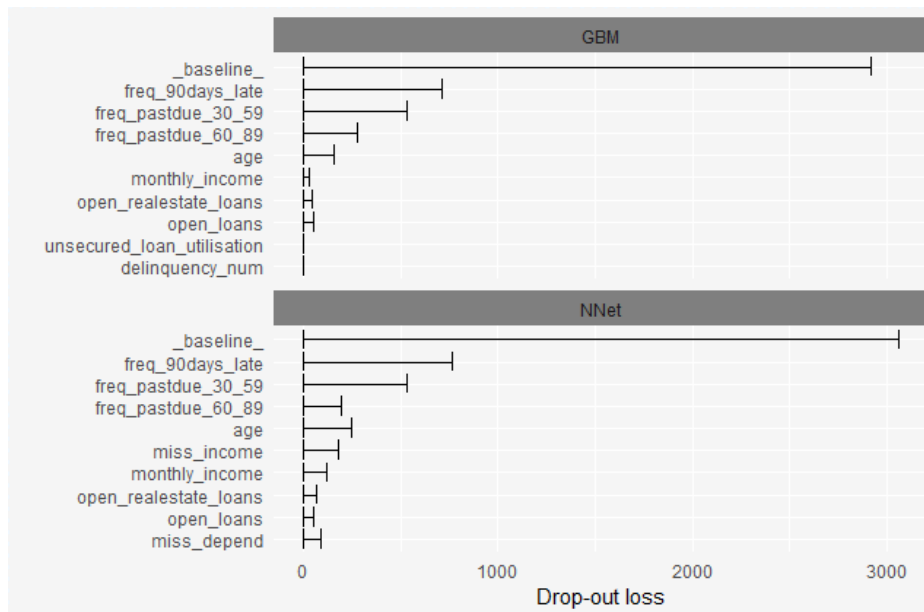
variable_importance()

A measure of the contribution of each variable to the model

Measured using **drop-out loss** (permutation loss):

- Randomly permute a predictor's values
- Score the model and measure the change in loss
- Repeat for each predictor

```
vi1 <- variable_importance(gbm_explain, loss_fn)
vi2 <- variable_importance(net_explain, loss_fn)
plot(vi1, vi2)
```





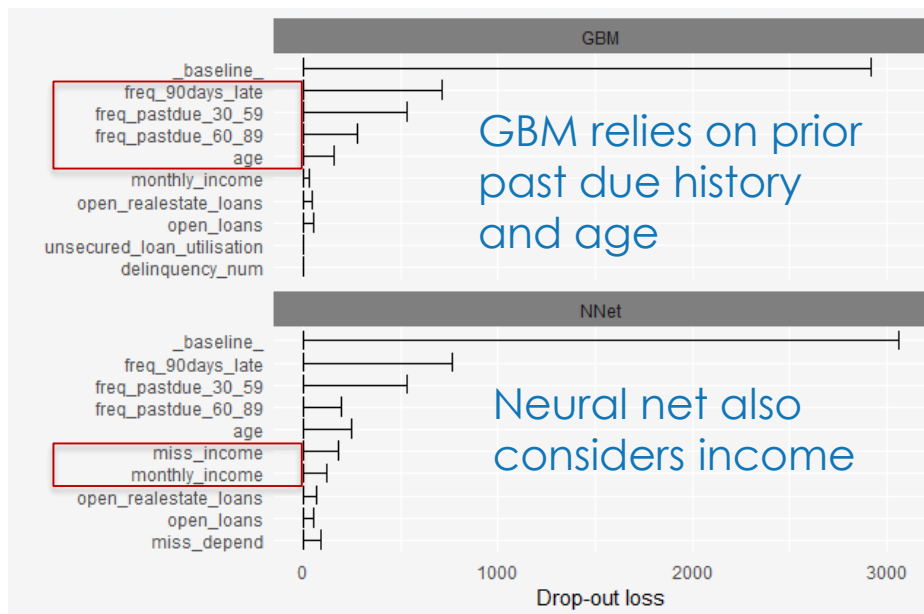
variable_importance()

A measure of the contribution of each variable to the model

Measured using **drop-out loss** (permutation loss):

- Randomly permute a predictor's values
- Score the model and measure the change in loss
- Repeat for each predictor

```
vi1 <- variable_importance(gbm_explain, loss_fn)
vi2 <- variable_importance(net_explain, loss_fn)
plot(vi1, vi2)
```





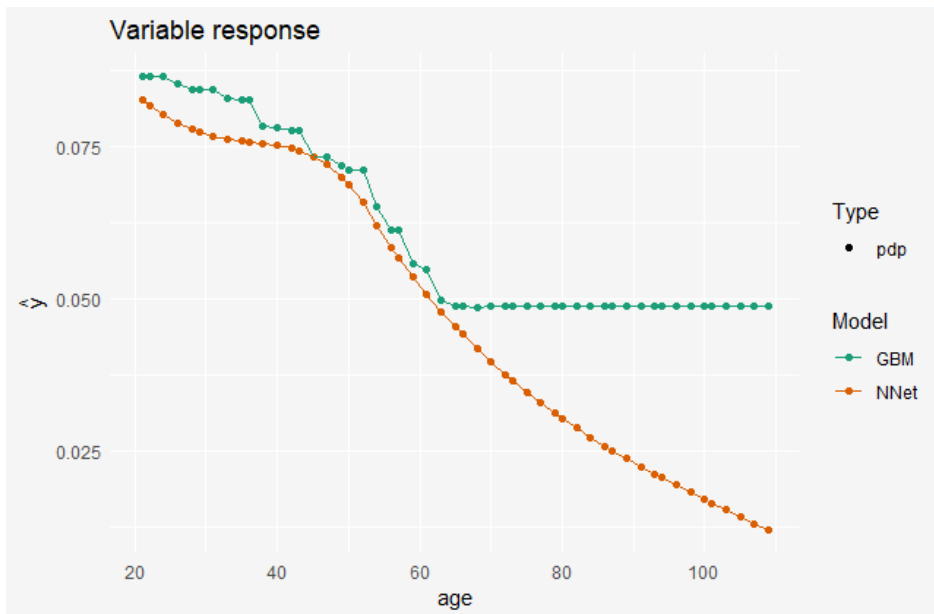
variable_response()

Plots the relationship between the prediction and the chosen variable(s)

Implements

- Partial Dependence Plots (shown right)
- Accumulated Local Effect plots (numeric only)
- Merging Path plots (factors only)

```
pd1 <- variable_response(gbm_explain, "age", "pdp")  
pd2 <- variable_response(net_explain, "age", "pdp")  
plot(pd1, pd2)
```





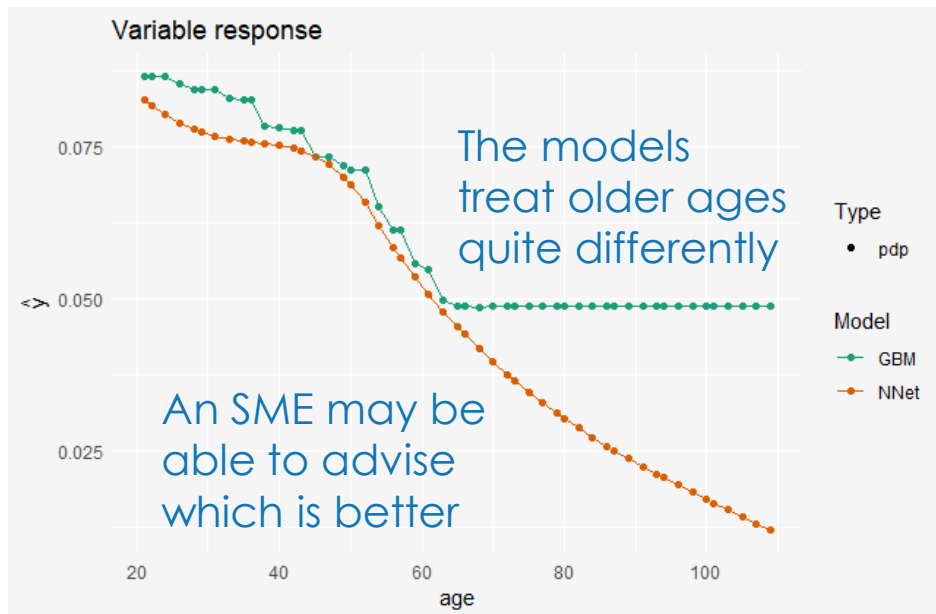
variable_response()

Plots the relationship between the prediction and the chosen variable(s)

Implements

- Partial Dependence Plots (shown right)
- Accumulated Local Effect plots (numeric only)
- Merging Path plots (factor only)

```
pd1 <- variable_response(gbm_explain, "age", "pdp")  
pd2 <- variable_response(net_explain, "age", "pdp")  
plot(pd1, pd2)
```





Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



Explaining individual predictions



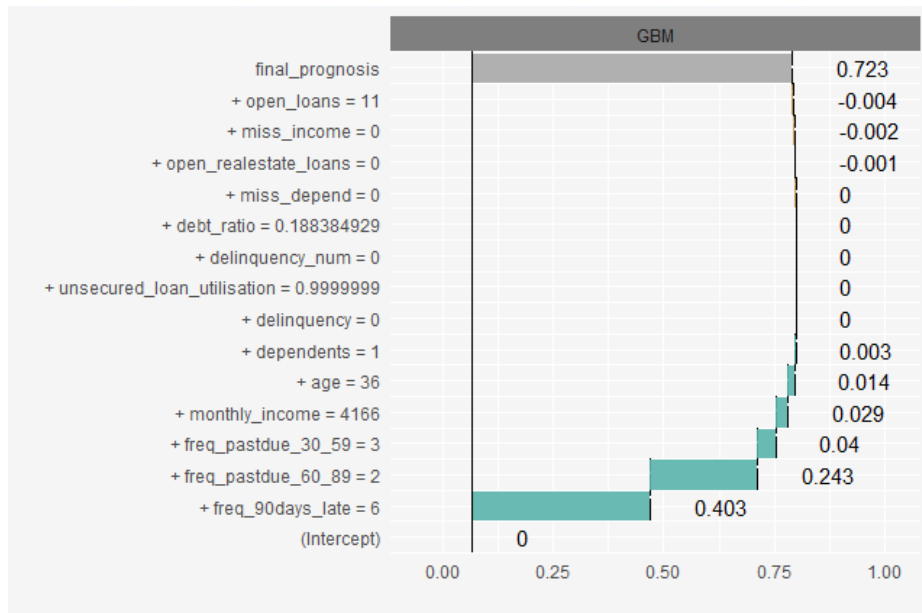
prediction_breakdown()

Summary of the **contribution of each factor** to the final prediction

Greedy calculation:

- Variables are removed from model one at a time, and deviation in prediction recorded
- Variable removed at each step is that which gives the smallest incremental deviation

```
obs <- data[row_number, ]  
p1 <- prediction_breakdown(gbm_explain, obs)  
plot(p1)
```





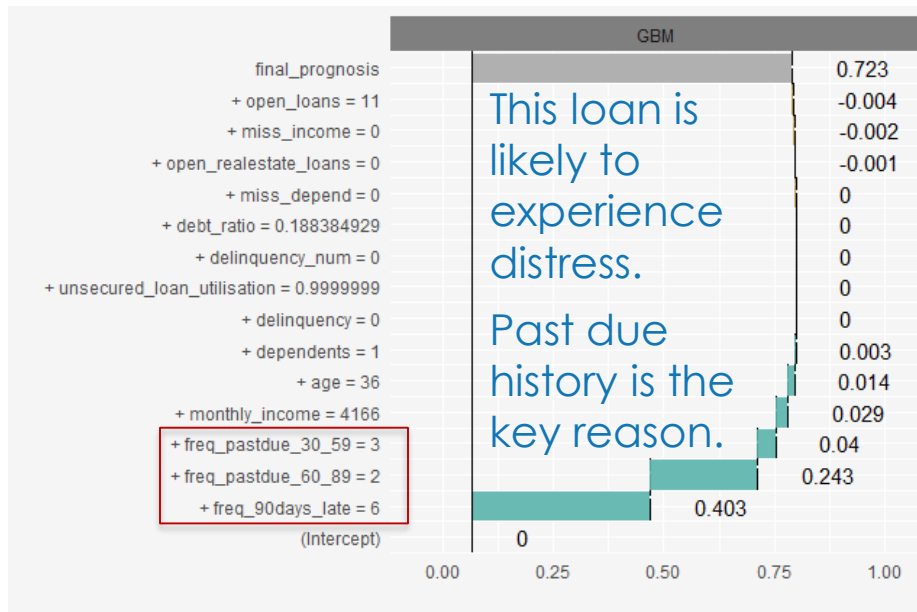
prediction_breakdown()

Summary of the **contribution of each factor** to the final prediction

Greedy calculation:

- Variables are removed from model one at a time, and deviation in prediction recorded
- Variable removed at each step is that which gives the smallest incremental deviation

```
obs <- data[row_number, ]  
p1 <- prediction_breakdown(gbm_explain, obs)  
plot(p1)
```





A brief look at LIME

Local Interpretable Model-agnostic Explanations

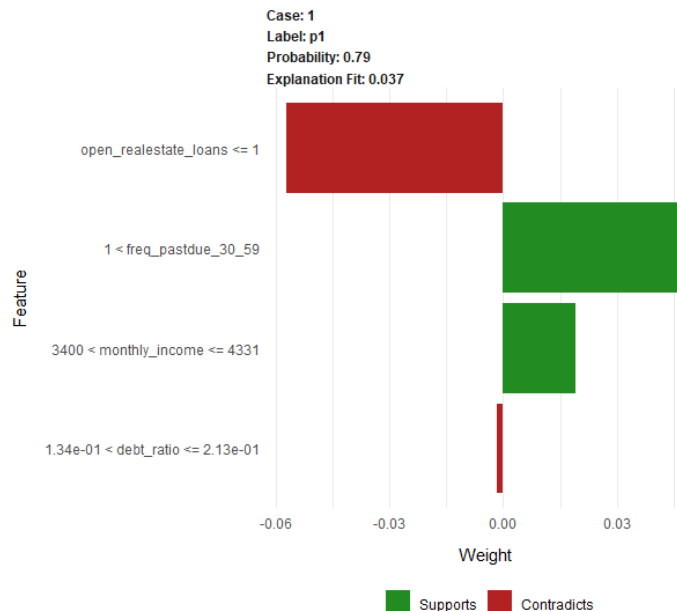
Provides interpretable explanations for the predictions of black-box classification models

Learns an interpretable model in the region locally around the prediction

Available for R and Python

github.com/marcotcr/lime

```
exp <- lime(data[, -c("target")], model.gbm)
val <- explain(obs, exp, n_labels = 1, n_features = 4)
plot_features(val)
```





A brief look at LIME

Local Interpretable Model-agnostic Explanations

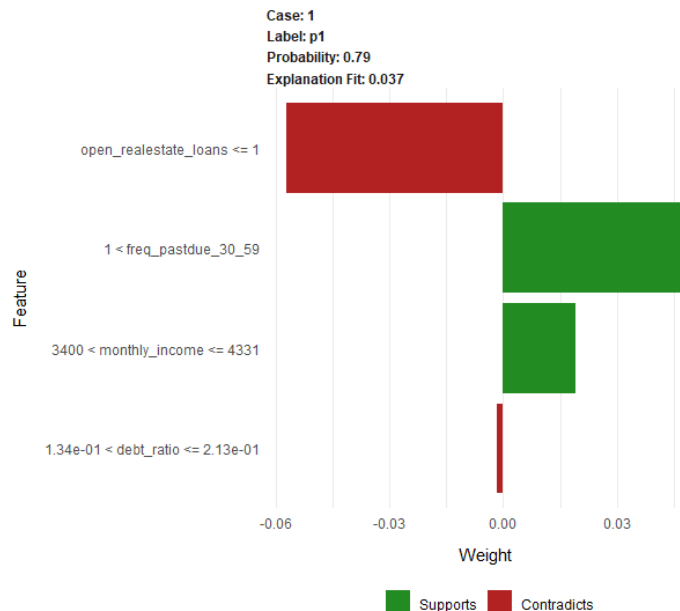
Provides interpretable explanations for the predictions of black-box classification models

Learns an interpretable model in the region locally around the prediction

Available for R and Python

github.com/marcotcr/lime

```
exp <- lime(data[, -c("target")], model.gbm)
val <- explain(obs, exp, n_labels = 1, n_features = 4)
plot_features(val)
```



History of frequent past due loans and a moderate income level support a higher likelihood of distress

A low number of open loans contradicts



Data Analytics Seminar

Reinvigorating the Usual

31 October 2018 • Actuaries Institute, Sydney



Thank you