Institute and Faculty of Actuaries

Gen Re

# Machine Learning Actuaries

Louis Rossouw, Gen Re

Chengdu IFoA Asia Conference 2019
9–10 May, Chengdu, China

## Overview of Machine Learning Techniques

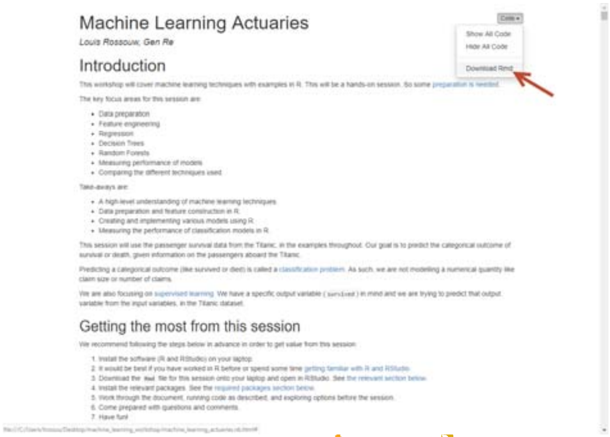
- Logistic Regression
- Decision Trees
- Random Forest
- Evaluation of Classification Models
- Other points to consider

## Workshop & Presentation

- Access R Notebook
  - Download with presentation

  OR

  - Download from RPubs
- Open in browser
- Follow instructions:
  - Download code
  - Install R & RStudio
- Learn to DIY in R!
- Slides follow R Notebook (broadly)

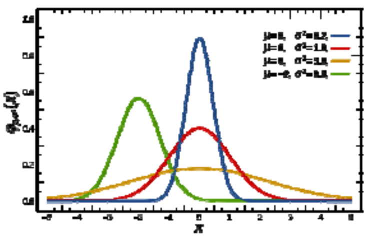



Institute and Faculty of Actuaries



## Statistical Learning vs. (pure) Machine Learning

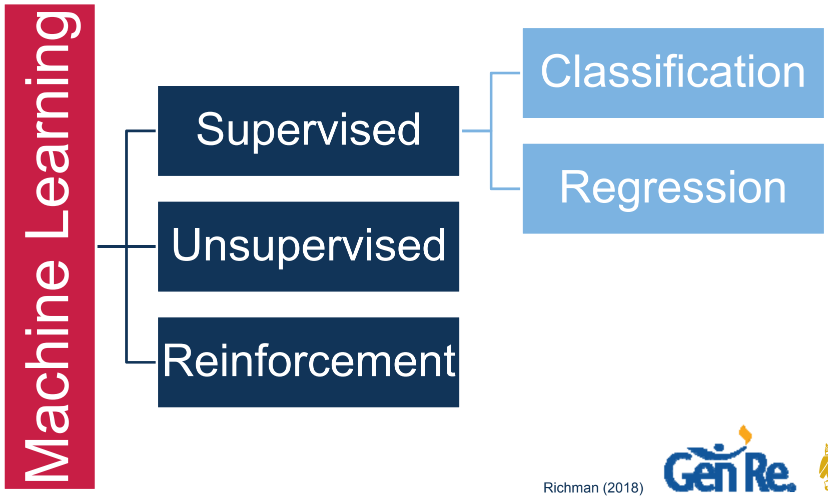
- Statistical / mathematical origins
- Statistical Models take account of uncertainty explicitly
- Structured (additive) predictor effects
- Can allow for complexity



- Programming / Computer Science origins
- Algorithmic with no predefined relationships
- Difficult to isolate effect of variables
- Easily deal with complexity

Institute and Faculty of Actuaries

## Machine Learning Overview




Richman (2018)



## Use cases for classification problems

- Predict a decision
  - Underwriting decision (accept at standard – Y/N)
  - Credit decision
- Propensity modelling
  - Propensity to lapse on month to month
  - Propensity to buy
- Mortality
  - Though often Poisson regression is more convenient (exposure)

## Titanic Survivor Data

- Passenger List of the Titanic
- Survival indicator
- Categorical outcome
- Split between training (75%) and testing data (25%)

| Field | Description |
|---|---|
| **pclass** | Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd) |
| **survival** | Survival (0 = No; 1 = Yes) |
| **name** | Name |
| **sex** | Sex |
| **age** | Age |
| **sibsp** | Number of Siblings/Spouses Aboard |
| **parch** | Number of Parents/Children Aboard |
| **ticket** | Ticket Number |
| **fare** | Passenger Fare |
| **cabin** | Cabin |
| **embarked** | Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton) |
| **home.dest** | Home/Destination |

Institute and Faculty of Actuaries



## Logistic Regression

- Bernoulli Distribution
- $logit(p) = ln\frac{p}{1-p}$
- $logit(p) = \sum x_i\beta_i$
- $p = \frac{1}{1+e^{\sum x_i\beta_i}}$
- $odds = e^{\sum x_i\beta_i}$
- $odds\ ratio = e^{\beta_i}$

- The odds are multiplied by $e^{\beta_i}$ for every unit increase in $x_i$
- If $x_i$ is an indicator (1 or 0) then $e^{\beta_i}$ is simply the odds ratio the event given data point is in that class (relative to not being in that class)

Institute and Faculty of Actuaries

## Interpretation of parameters

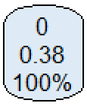
- Predicting survival
- Odds ratio for age $e^{-0.010089} = 0.990$
- I.e. odds of survival decrease by 1% for every year increase in age
- Odds ratio for a Miss $e^{0.216780} = 1.242$
- I.e. odds of a "Miss" survival is 24.2% higher than a "Master" surviving.

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.945421   0.486424   3.999 6.35e-05 ***
titleMiss      0.216780   0.410777   0.528  0.59768
titleMr       -2.605564   0.432944  -6.018 1.76e-09 ***
titleMrs       0.681932   0.449904   1.516  0.12959
titleOfficial -1.835108   0.683201  -2.686  0.00723 **
family_size   -0.432111   0.073558  -5.874 4.24e-09 ***
embarkedQ     -0.907650   0.344113  -2.638  0.00835 **
embarkedS     -0.541077   0.215449  -2.511  0.01203 *
age           -0.010089   0.007967  -1.266  0.20539
fare           0.011518   0.002338   4.926 8.40e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gen Re. Institute and Faculty of Actuaries

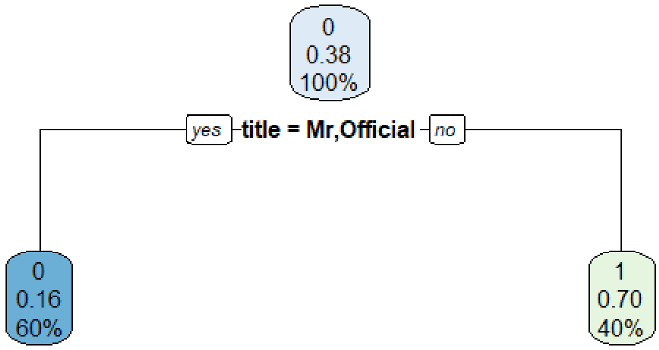


## The data




Institute and Faculty of Actuaries

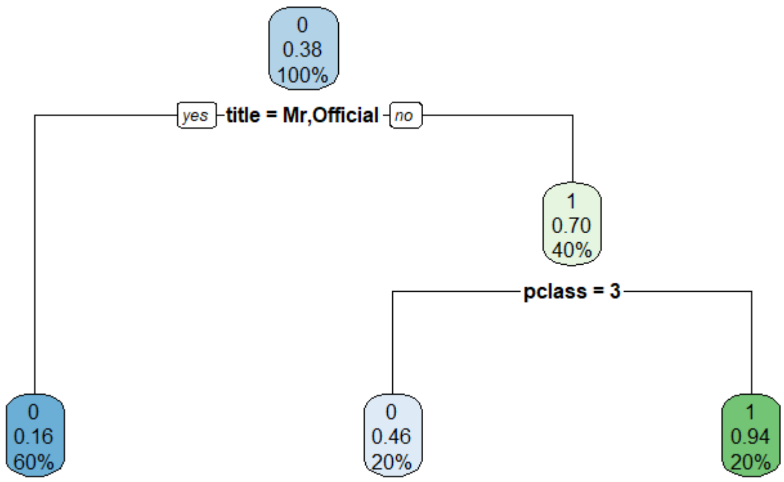
# Decision Trees – Depth 1




Institute and Faculty of Actuaries

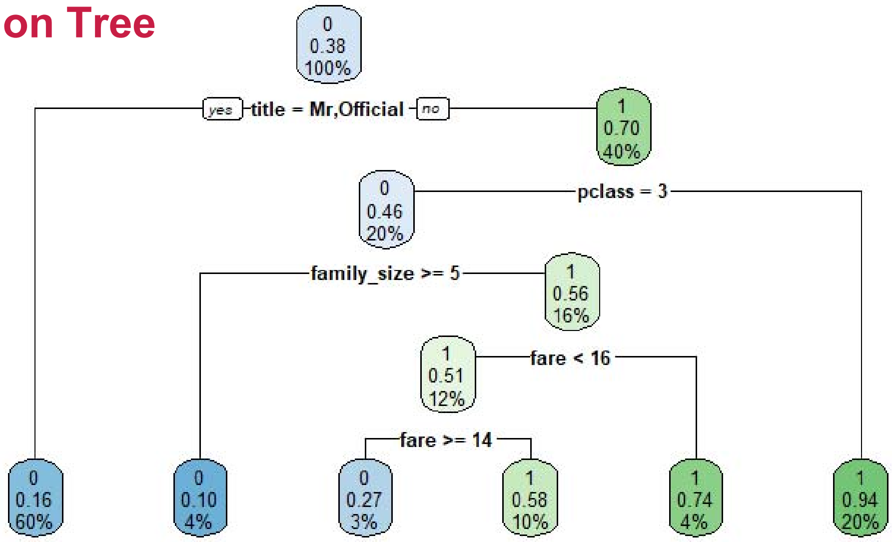


# Decision Tree – Depth 2




Institute and Faculty of Actuaries

## Full Decision Tree






## Other points on decision trees

- Predictions are made based on the observed probabilities in the leaf nodes
  - If p>0,5 = predict survival
  - Or we can simply use the probability as a score
- In the above example Gini impurity was used to decide best splits
- Various stopping conditions can be used
  - Impacts over- or underfitting
- Can interpret results (if tree remains small)

Institute and Faculty of Actuaries

## Ensemble Models

- Models where predictions from multiple models are combined
- We could combine different kinds of models
- But we could also combine many combinations of the same model
- **Forest** = many decision tree models
- Each tree is fit on a **random** subset of rows and columns

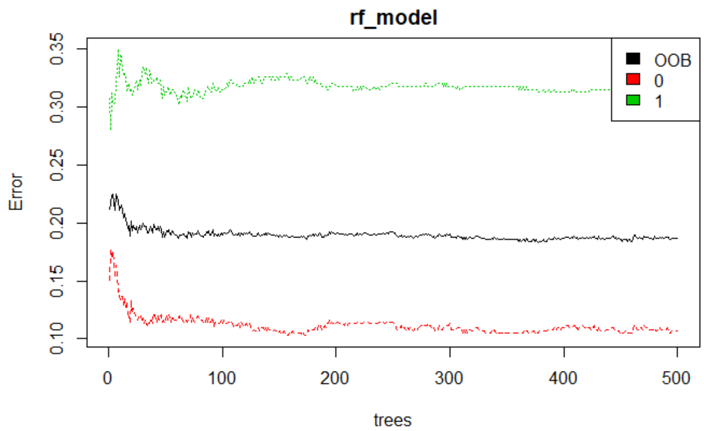
➢ **Random Forest**

- Prediction is based on aggregate prediction from trees
- 1 tree = 1 vote



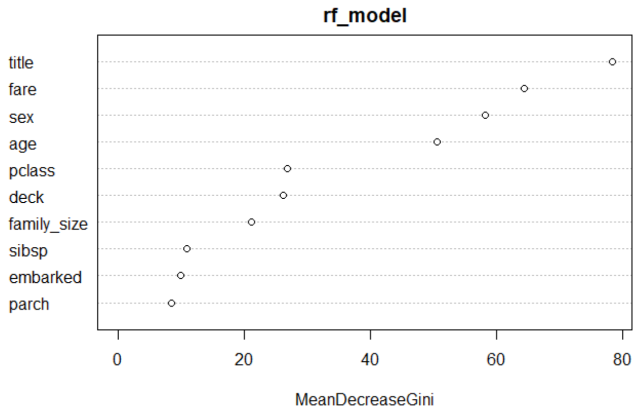
## Out of Bag Error Rates

- Each tree has data it was not trained on
- Calculate the error rate of the tree on the data it was not trained on
- Aggregate these error rates



## Interpretation is problematic…

- How do you review the impact of each variable?
- Same variable could be used multiple times in the same tree or different tree
- We have 500 trees…
- **Variable importance plot**
  - Sum the reduction in “impurity” every time a variable is used
  - Compare variables






Institute and Faculty of Actuaries

## Classification Model Evaluation

- Confusion Matrix
- Receiver operator characteristic
- Area under the curve
- Over- and underfitting

Rossouw (2018)

## What do we have?

| row | predict_prob_glm | predict_glm | survived |
|---|---|---|---|
| **1045** | 0.6680738 | 1 | 1 |
| **986** | 0.1426785 | 0 | 1 |
| **512** | 0.0748915 | 0 | 0 |
| **447** | 0.4889538 | 0 | 1 |
| **472** | 0.7950490 | 1 | 1 |
| **259** | 0.8562478 | 1 | 1 |

Institute and Faculty of Actuaries



## Confusion Matrix

| | Actual = 0 | Actual = 1 | Total |
|---|---|---|---|
| **Predicted = 0** | 171 | 39 | **210** |
| **Predicted = 1** | 27 | 90 | **117** |
| **Total** | **198** | **129** | **327** |

- **Accuracy** = (90 + 171) / 327 = **79.8%**
- **Sensitivity** = True Positive Rate = 90 / 129 = **69.8%**
- **Specificity** = True Negative Rate = 171 / 198 = **86.4%**
- This uses threshold p of 0.5

Institute and Faculty of Actuaries

## Change the threshold?

Predict survival if p>0.1

- Sensitivity = 98.4%
- Specificity = 14.6%
- Accuracy = 47.7%

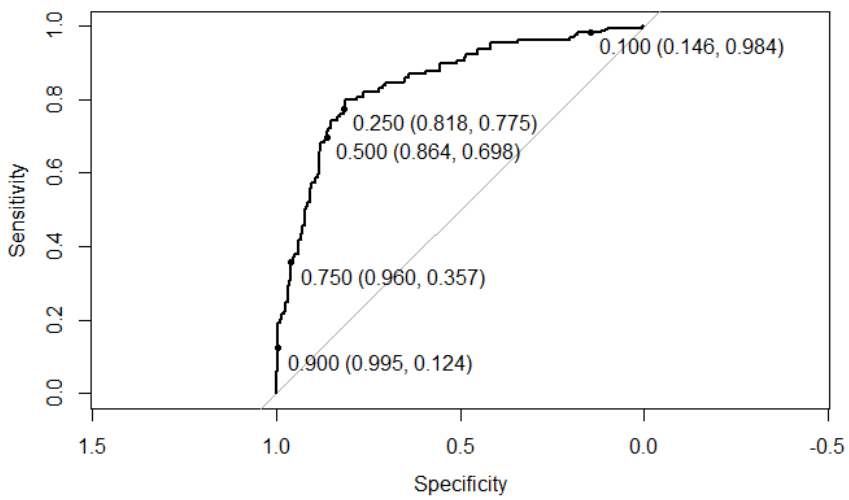
Predict survival if p>0.9

- Sensitivity = 12.4%
- Specificity = 99.5%
- Accuracy = 65.1%

Institute and Faculty of Actuaries



## Receiver Operator Characteristic (ROC) Curve for GLM




Institute and Faculty of Actuaries

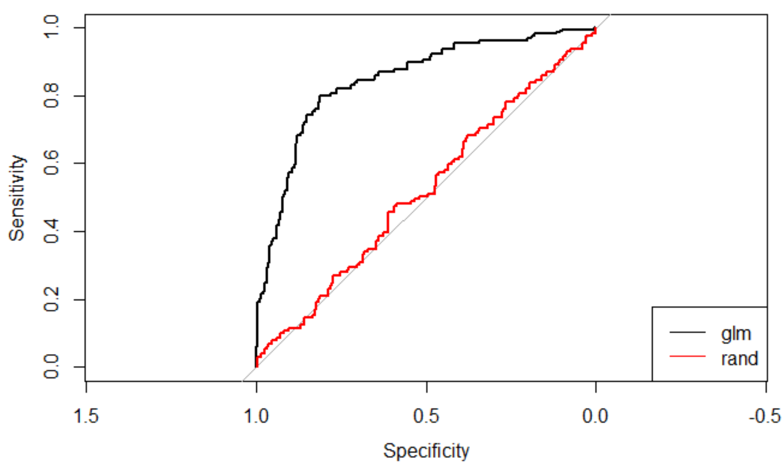
## Area Under the Curve (AUC)

- AUC is measure of overall performance of the model
- AUC = Probability that score of a random survivor > score of random person who died
- Gini Coefficient = 2 * AUC – 1
- AUC > 70% OK
- AUC > 80% good

Institute and Faculty of Actuaries



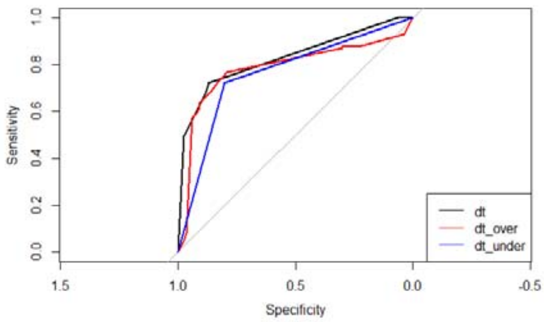
## Random Guessing




Institute and Faculty of Actuaries

## Over- vs. underfitting

| Model | Training AUC | Testing AUC |
|---|---|---|
| Decision Tree – Underfitted | 77.5% | 76.1% |
| Decision Tree | 83.3% | 82.7% |
| Decision Tree – Overfitted | 91.5% | 78.3% |




Institute and Faculty of Actuaries

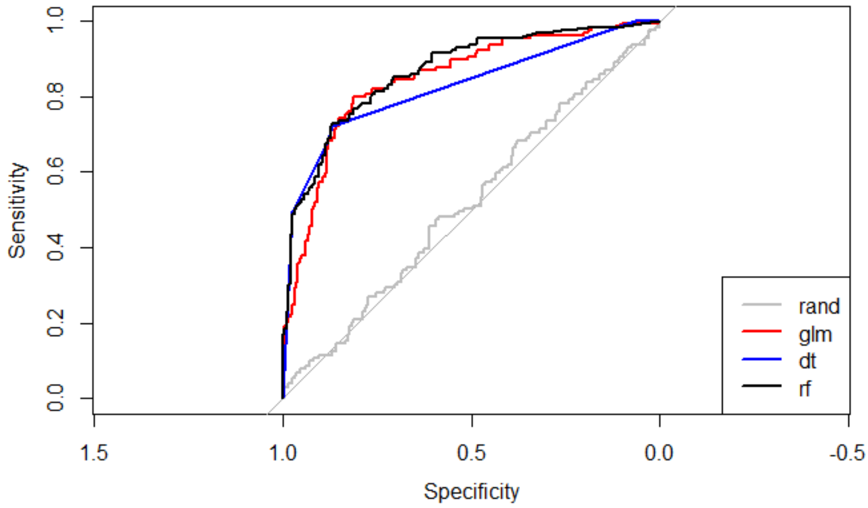


## Model Comparison

| Model | AUC |
|---|---|
| **Random Guessing** | 51.9% |
| **GLM** | 84.6% |
| **Decision Tree** | 82.7% |
| **Random Forest** | 86.7% |

Institute and Faculty of Actuaries

## Model Comparison – ROC




Institute and Faculty of Actuaries



## Shortcomings of ROC / AUC

- Measures classification
- The probabilities are not calibrated
- Random Forest does not strictly produce a probability
  - a proportion of votes of trees
- Measures the accuracy of “ordering” of data

historiska

Institute and Faculty of Actuaries

## Other considerations when deciding on a model

- How important is interpretability?
  – Do you need to be able to explain the model in detail?
- Technical issues
  – Computation speed & resources
- Do you need to be able to explain the model in depth?

Institute and Faculty of Actuaries



Institute and Faculty of Actuaries

## Further thoughts

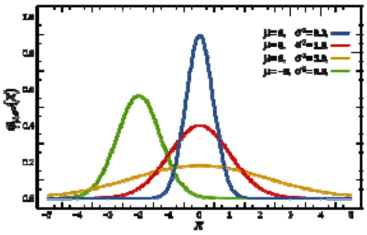
- Other machine learning techniques
- Opening the black box
- Testing data
- Cross-validation
- Hyperparameter tuning

Rossouw (2018)

## Other Machine Learning Algorithms

Statistical Learning

- Generalised Linear Regression
- Generalised Additive Models
- Penalised Regression
- …



Machine Learning

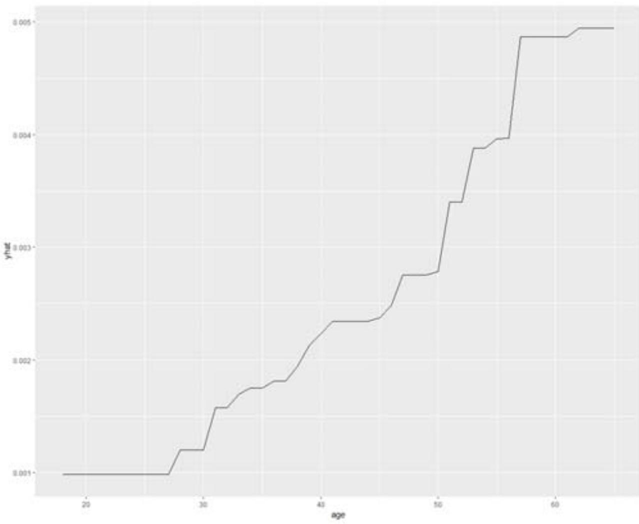
- Gradient Boosted Machines
- Support Vector Machines
- (Deep) Neural Networks
- …

Institute and Faculty of Actuaries



## Opening the black box…

- Variable Importance Plot
- Partial Dependence Plot
- Surrogate model
  - Simple decision tree
  - Local interpretable model-agnostic explanations (LIME)




Jalali (2018)

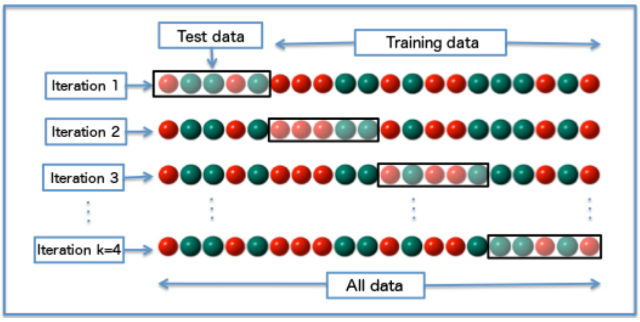
# Testing (hold-out) Data

- Statistical models
  - Test validity of the model using statistics
  - Hold-out data is not required (but can be good)
- Machine Learning Models
  - Maybe prone to overfit etc.
  - Hold-out data validates that it did not occur
  - Hold-out should not be used repeatedly to refine model
- Also consider cross-validation

Institute and Faculty of Actuaries



# Cross-validation




Fabian Flöck

1. Split data into k datasets
   - Called folds (e.g. 4)
   - 4 separate datasets
2. Fit model on 3 folds
3. Calculate metric (e.g. AUC/error rate) on remaining fold
4. Repeat 4 times until each fold has been held back
5. Average/aggregate error metrics across the 4 folds

Institute and Faculty of Actuaries

## Hyperparameter tuning

- ML techniques require many parameters
  - Maximum depth
  - Minimum child weights
  - Number of variables selected
  - Number of data rows selected
  - …
- Search parameters that minimise error / maximise accuracy
  - Grid / random / ranges
  - Cross validation
- Still validate with a hold-out dataset if possible

Gen Re. Institute and Faculty of Actuaries



## Conclusion

- Overview of machine learning techniques
  - Logistic regression
  - Decision trees
  - Random Forest
- Evaluation of classification models
  - Confusion matrix
  - ROC curve & AUC
  - Other considerations
- Further thoughts

Gen Re. Institute and Faculty of Actuaries

## We did not cover

- Data validation
- Feature engineering
- Regression problems
  - Poisson
- Ensemble techniques
- And so much more…

Institute and Faculty of Actuaries



## References


- Richman (2018), AI in Actuarial Science.
- Rossouw (2018), Classification Model Performance
- Jalali (2018), Unveiling Black Box Models – Interpretability and Trust


Institute and Faculty of Actuaries

Questions

Comments

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter.

Institute and Faculty of Actuaries