Institute and Faculty of Actuaries

Fraud Detection: How can Machine Learning Help?

Navarun Jain
Lux Actuaries & Consultants
Dubai

02 September 2019

Agenda

Insurance Fraud – the Issue

ML as a Concept

Tree-based ML Algorithms

- CART
- C5.0
- Gradient Boosting Machines
- Random Forests

Non-tree-based ML Algorithms

- Neural Networks

Applications to Fraud Detection – A Simple Case Study

- Data Description
- Models and Performance Evaluation
- Interpretation

Key Takeaways and Conclusions

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Insurance Fraud

02 September 2019

Insurance Fraud

Institute and Faculty of Actuaries

Insurance Fraud

- Becoming a rapidly growing issue worldwide

Institute and Faculty of Actuaries

Insurance Fraud

- Becoming a rapidly growing issue worldwide
- UK fraud activity reached an estimated £17 million in 2018

Institute and Faculty of Actuaries

Insurance Fraud

- Becoming a rapidly growing issue worldwide
- UK fraud activity reached an estimated £17 million in 2018
- Biggest lines are Motor, Medical and Workmen’s Compensation – fake car crashes, personal injury scams, faked death claims

Institute and Faculty of Actuaries

Insurance Fraud

- Becoming a rapidly growing issue worldwide
- UK fraud activity reached an estimated £17 million in 2018
- Biggest lines are Motor, Medical and Workmen’s Compensation – fake car crashes, personal injury scams, faked death claims
- With advancing technology, it can become easier to detect fraudulent claims when they are received

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Machine Learning

02 September 2019

Machine Learning

The ”teaching a kid math” analogy

The Roadmap

All about patterns!!!

Computer systems learn
from data

We **train** the system → System **learns** → Then performs operations **on its own**

Institute and Faculty of Actuaries

The Roadmap

All about patterns!!!

Computer systems <u>learn</u> from data

We **<u>train</u>** the system → System **<u>learns</u>** → Then performs operations **<u>on its own</u>**

Training phase 1: data is fed into the algorithm, relevant fields and records sorted from data to retrieve **active dataset**

Institute and Faculty of Actuaries

The Roadmap

All about patterns!!!

Computer systems <u>learn</u> from data

We **<u>train</u>** the system → System **<u>learns</u>** → Then performs operations **<u>on its own</u>**

Training phase 2: **Model Fitting** – algorithm decodes hidden patterns and relationships in the data

Institute and Faculty of Actuaries

The Roadmap

All about patterns!!!

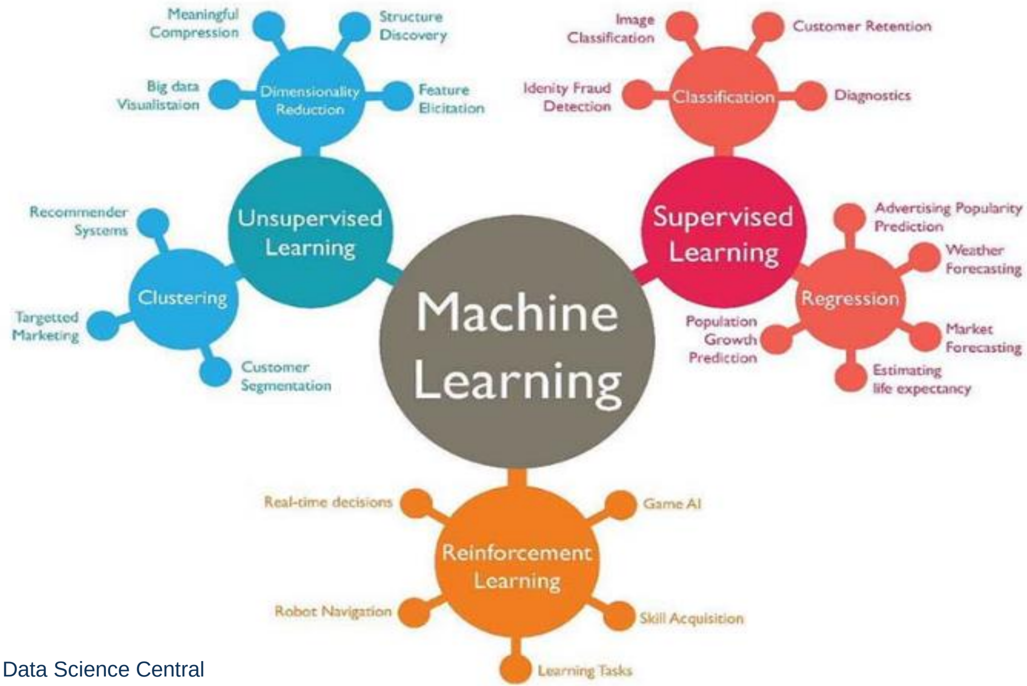
Computer systems learn from data

We **train** the system → System **learns** → Then performs operations **on its own**

Testing phase: new data fed into system, algorithm uses patterns & relationships learnt during the training phase to predict new cases

Institute and Faculty of Actuaries

Types of Algorithms



Source: Data Science Central

Institute and Faculty of Actuaries

With ML, no need to…

Institute and Faculty of Actuaries

With ML, no need to…

- …make assumptions about distributions
- …worry about possible correlations between predictors
- …look for interactions between predictors

Institute and Faculty of Actuaries

How can ML help?

RULE-BASED FRAUD DETECTION	ML-BASED FRAUD DETECTION
Can catch obvious and known fraud scenarios only	Can find not-so-obvious fraud scenarios due to the ability to detect hidden patterns/correlations in data
Requires manual work to determine criteria for fraud scenarios	Can automatically detect and create rules for fraud scenarios
Longer processing and verification times due to manual nature	Quicker processing and verification times since algorithms are automatically generated and verified

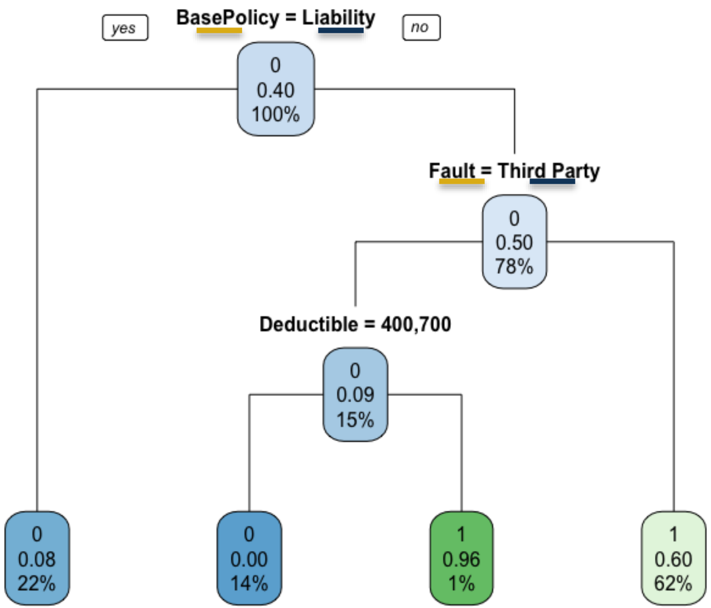
Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Tree-Based ML Algorithms

02 September 2019

Decision Trees



Model is grown by recursively splitting the data into **decision boundaries** using the **feature space**

Institute and Faculty of Actuaries

Types of Decision Tree Algorithms

SINGLE TREE MODELS

1. CART
2. C5.0

ENSEMBLE MODELS

1. GBM
2. RANDOM FOREST

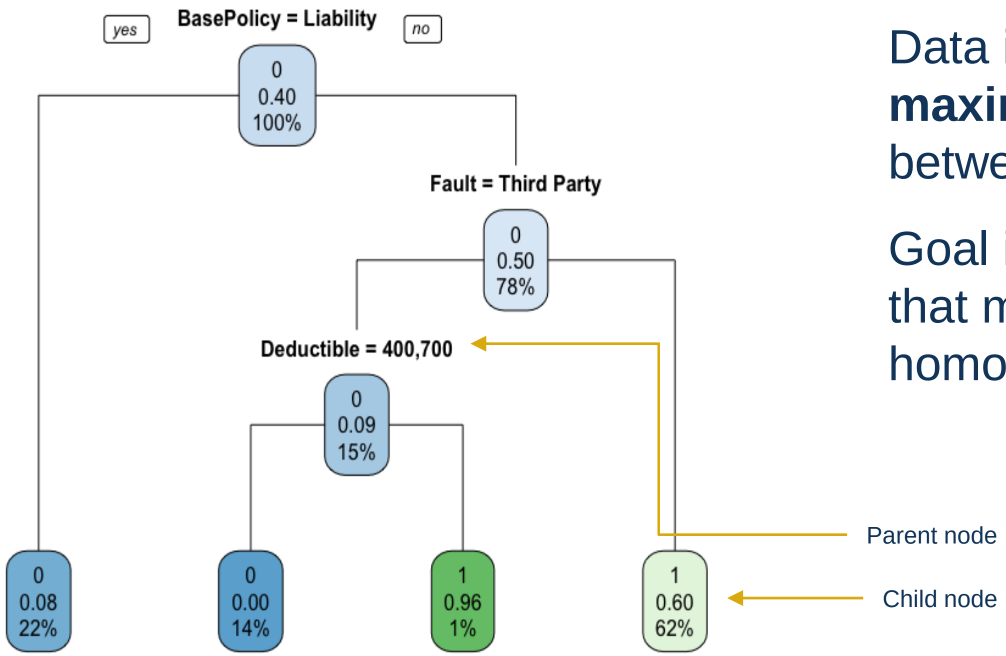
Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Single Tree Models

02 September 2019

Creating a Decision Tree



Data is split in a way that **maximizes** the gain in information between *parent* and *child* nodes

Goal is to split data points in a way that makes the subgroups as homogenous as possible

Institute and Faculty of Actuaries

Measuring Information Gain

Measuring Information Gain

Gini Impurity

$$Gini(t) = \sum_{k=1}^{h} p_k(1 - p_k)$$

p_k – Probability of choosing item with label k in set t

Measures how often a randomly chosen element would be incorrectly labeled if it were labeled according to its distribution in the data

Used as splitting criterion for the **CART** algorithm

Entropy

$$H(t) = -\sum_{k=1}^{h} \{p_k \log_b p_k\}$$

p_k – Probability of choosing item with label k in set t
b – Logarithmic base

Measures how "mixed up" the data is

Used as splitting criterion for the **C5.0** algorithm

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Ensemble Learning Models

02 September 2019

Institute and Faculty of Actuaries

Gradient Boosting Machines

02 September 2019

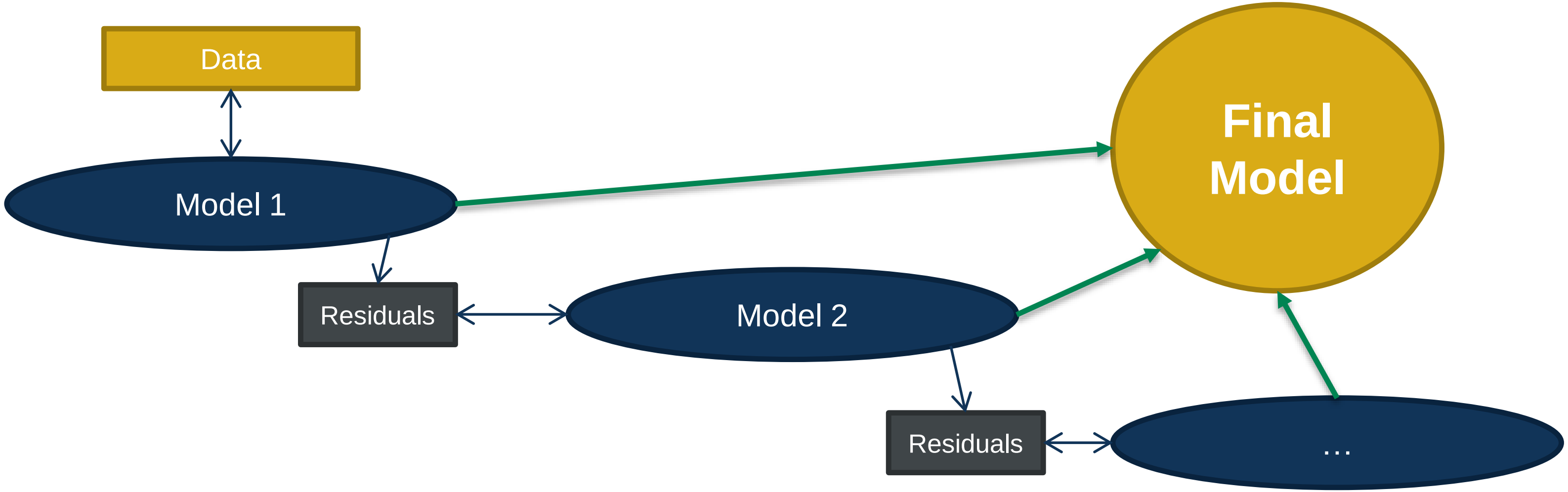
Boosting

- Converts weak learners into a single strong learner by aggregating them

Institute and Faculty of Actuaries

Boosting

- Converts weak learners into a single strong learner by aggregating them



Institute and Faculty of Actuaries

Random Forests

02 September 2019

Breaking Down the “Random Forest”

Breaking Down the “Random Forest”

- RF based on the concept of **Bagging** (**B**ootstrap **Agg**regat**ing**)

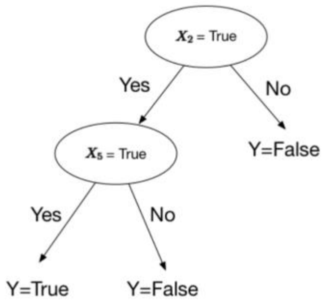
Breaking Down the "Random Forest"

- RF based on the concept of **Bagging** (**B**ootstrap **Agg**regat**ing**)

Transaction	Time stamp	ID	Age group	Fresh fruit	Seafood
1	April	1	Senior	Orange	Tuna
2	April	2	Teenager	Banana	Tuna
3	April	4	Teenager	Banana	Tuna
4	June	1	Senior	Orange	Anchovy
5	June	3	Adult	Banana	Tuna
6	July	1	Senior	Orange	Anchovy
7	July	2	Teenager	Banana	Tuna
8	July	2	Teenager	Orange	Tuna
9	July	4	Teenager	Banana	Tuna
10	December	2	Teenager	Banana	Tuna
11	December	3	Adult	Orange	Anchovy
12	December	1	Senior	Orange	Anchovy

Random sample of *p* columns

Random sample of *k* rows



Repeat *N* times

Institute and Faculty of Actuaries

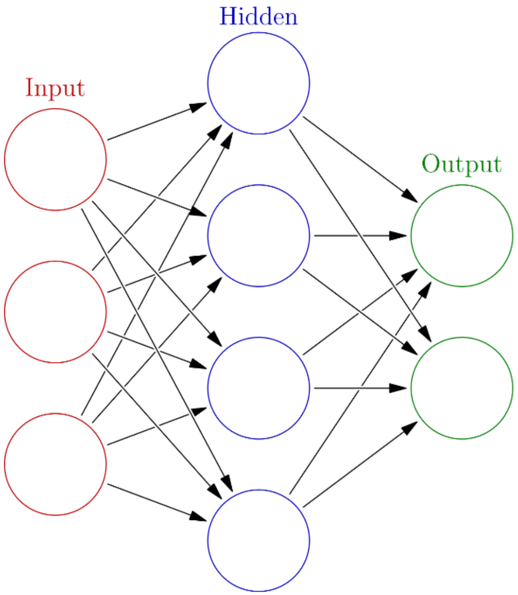
Institute and Faculty of Actuaries

Artificial Neural Networks

02 September 2019

Artificial Neural Networks

Structured Sequential model



Structured: A Neural Network has a defined structure that consists of 3 types of layers

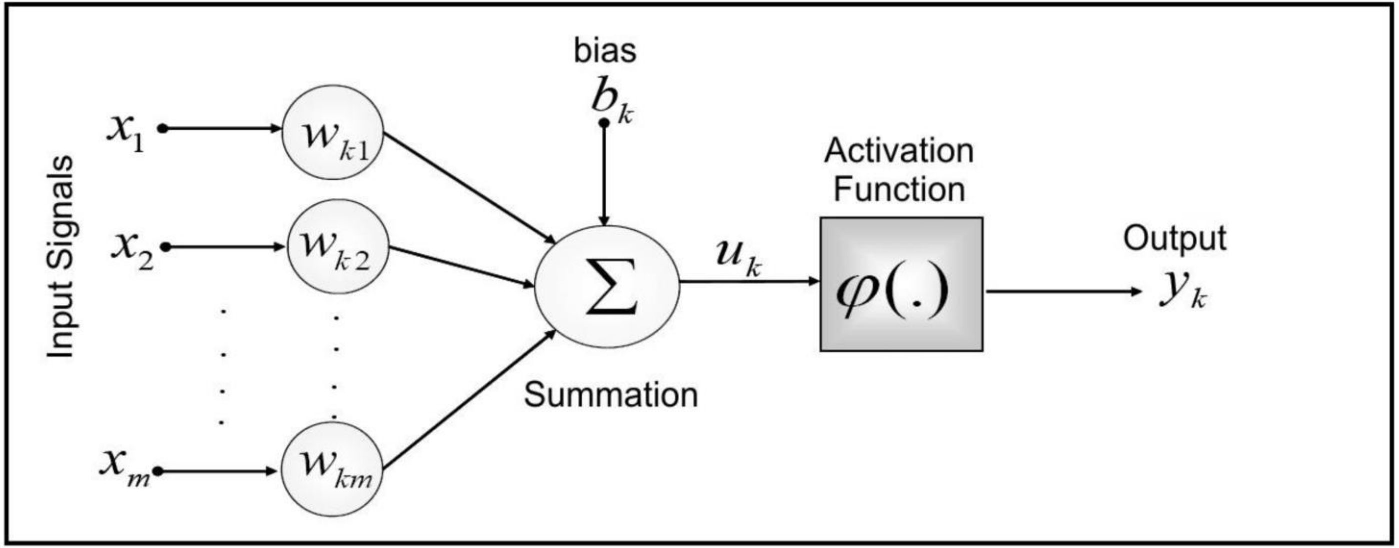
Sequential: Information flows in a sequence from one layer to the next, undergoing operations at each layer – almost like an assembly line

Institute and Faculty of Actuaries

How ANN's Work

Institute and Faculty of Actuaries

How ANN's Work



Institute and Faculty of Actuaries

How ANN's Work

- Data in every neuron is transformed by an activation function:

$$h_k(x) = g(\beta_{0k} + \sum_{i=1}^{n} x_i \beta_{ik})$$

$h_k(x)$ – k^{th} neuron in a hidden layer
β_{ik} - coefficient of the i^{th} previous-layer neuron on above neuron

Institute and Faculty of Actuaries

How ANN's Work

- Data in every neuron is transformed by an activation function:

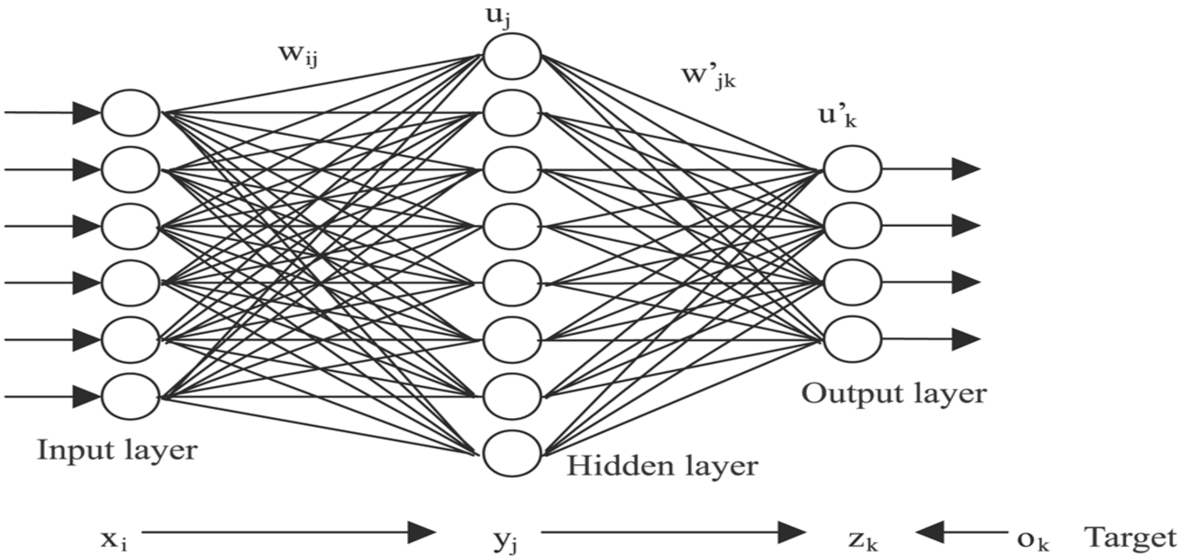
$$h_k(x) = g(\beta_{0k} + \sum_{i=1}^{n} x_i \beta_{ik})$$

$h_k(x)$ – k^{th} neuron in a hidden layer
β_{ik} - coefficient of the i^{th} previous-layer neuron on above neuron

- Activation function transforms the linear combination of inputs from one layer and sends it to the next layer.

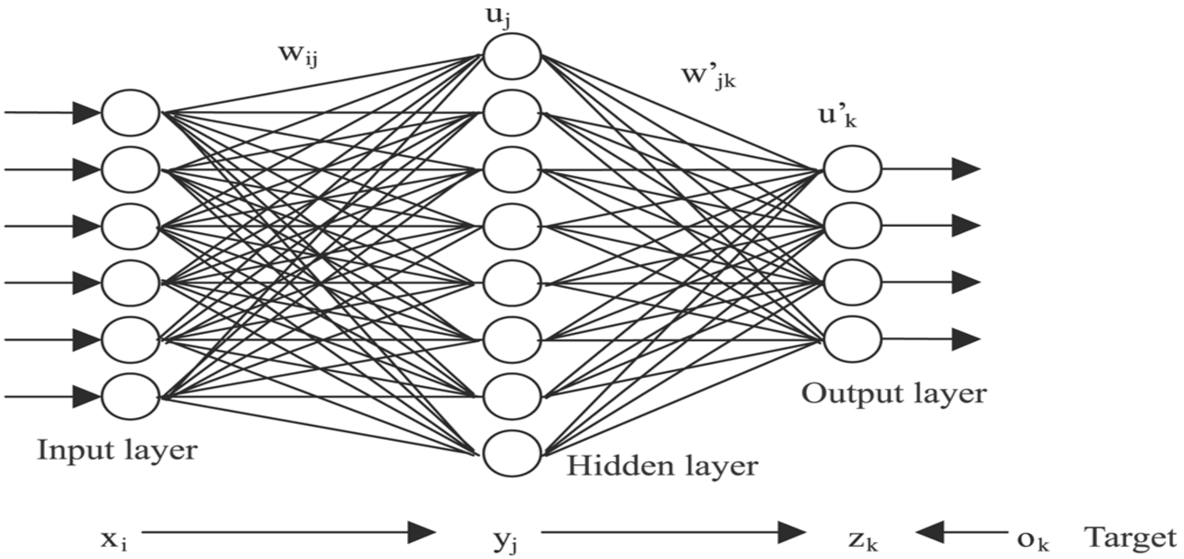
Institute and Faculty of Actuaries

How ANN's Work



Institute and Faculty of Actuaries

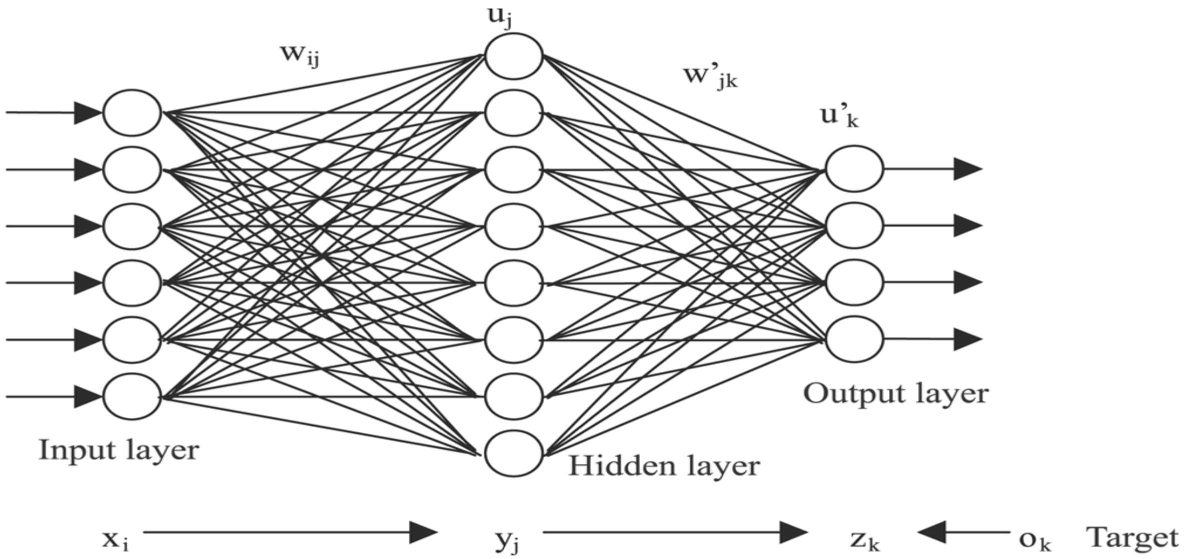
How ANN’s Work



At first, each neuron is randomly assigned a weight – this measures the contribution of that neuron to the next layer

Institute and Faculty of Actuaries

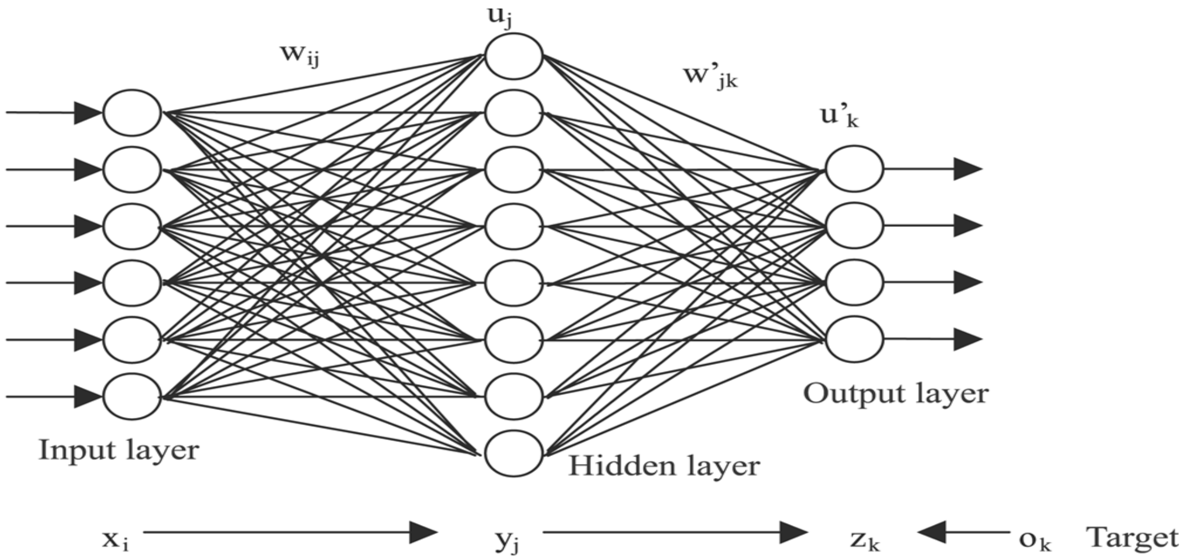
How ANN's Work



Data flows through network, predicted values calculated

Institute and Faculty of Actuaries

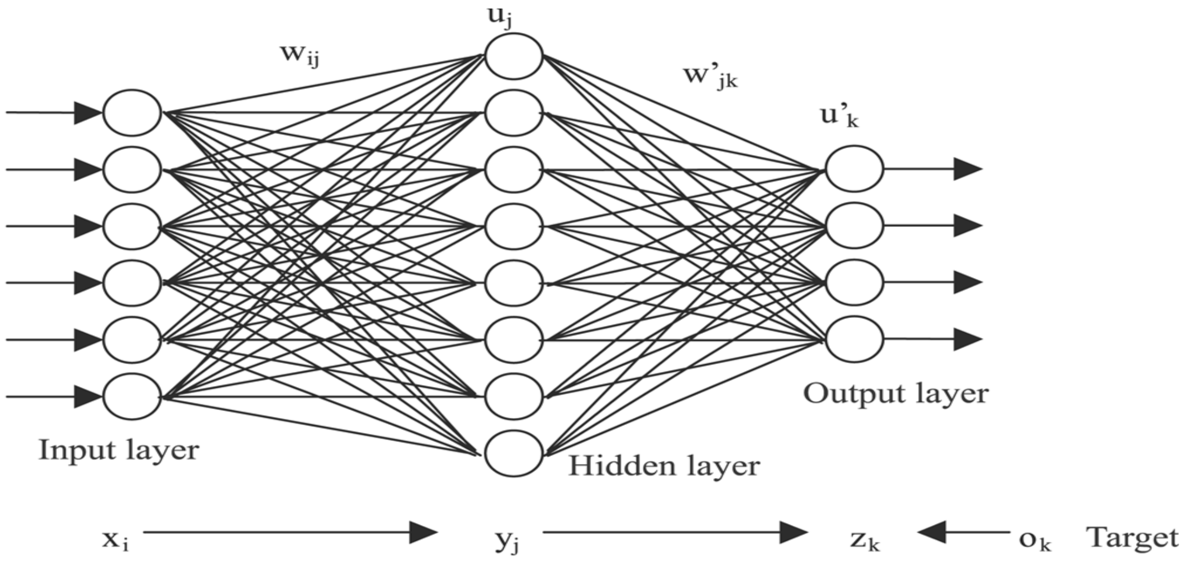
How ANN's Work



Predictions are compared with actuals based on a loss function

Institute and Faculty of Actuaries

How ANN’s Work



Weights are updated to reduce value of loss function

Institute and Faculty of Actuaries

Institute
and Faculty
of Actuaries

Case Study: Classifying Motor Insurance Fraud

02 September 2019

The Data

- Claim-level information with an indicator for whether a claim was flagged as a fraud or not
- Data points for each claim include –
  - Driver demographics (age, marital status, gender)
  - Vehicle information (age, price, body type, country or origin)
  - Policy information (policy cover type, number of vehicles insured, deductible, agent type)
  - Accident/Claim information (when was the claim filed, whether there were witnesses present during the accident, party at fault, whether a police report was filed)

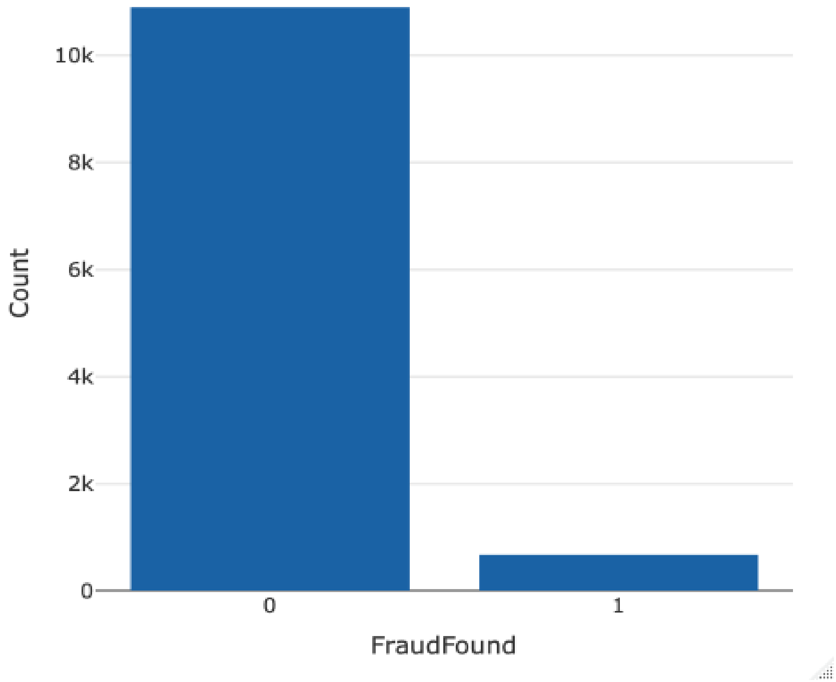
Institute and Faculty of Actuaries

Summary of Results

- GBM, Random Forest performed best, followed by Neural Networks
- C5.0, CART poor
- Logistic Regression did not perform well
- Driver Age, Policy Type, Fault, Past Number of Claims most important predictors of fraudulent behavior
- Details in following slides

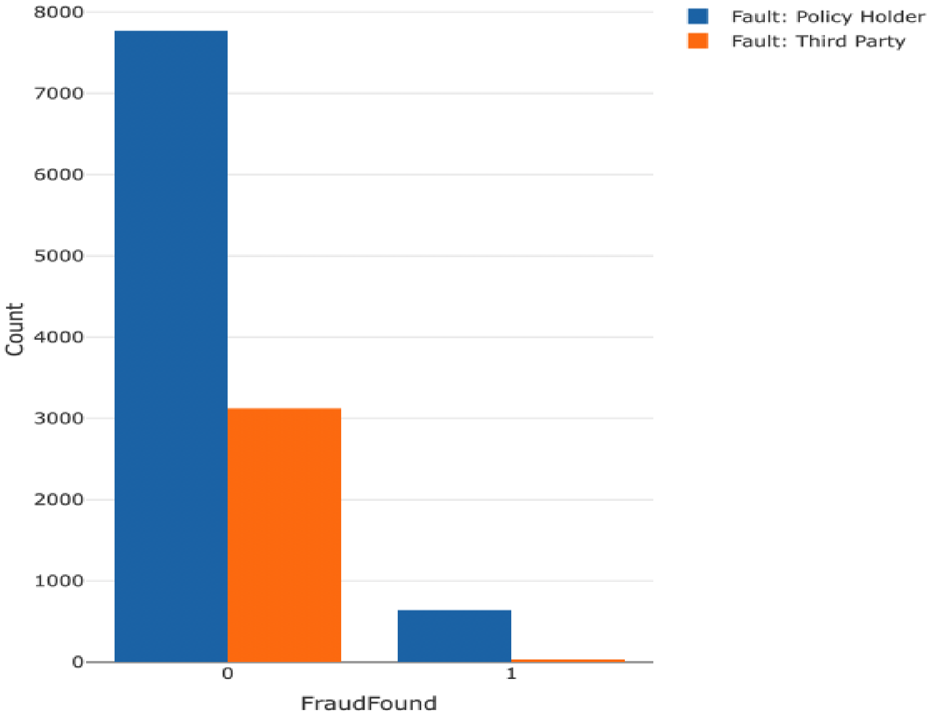
Institute and Faculty of Actuaries

The Data



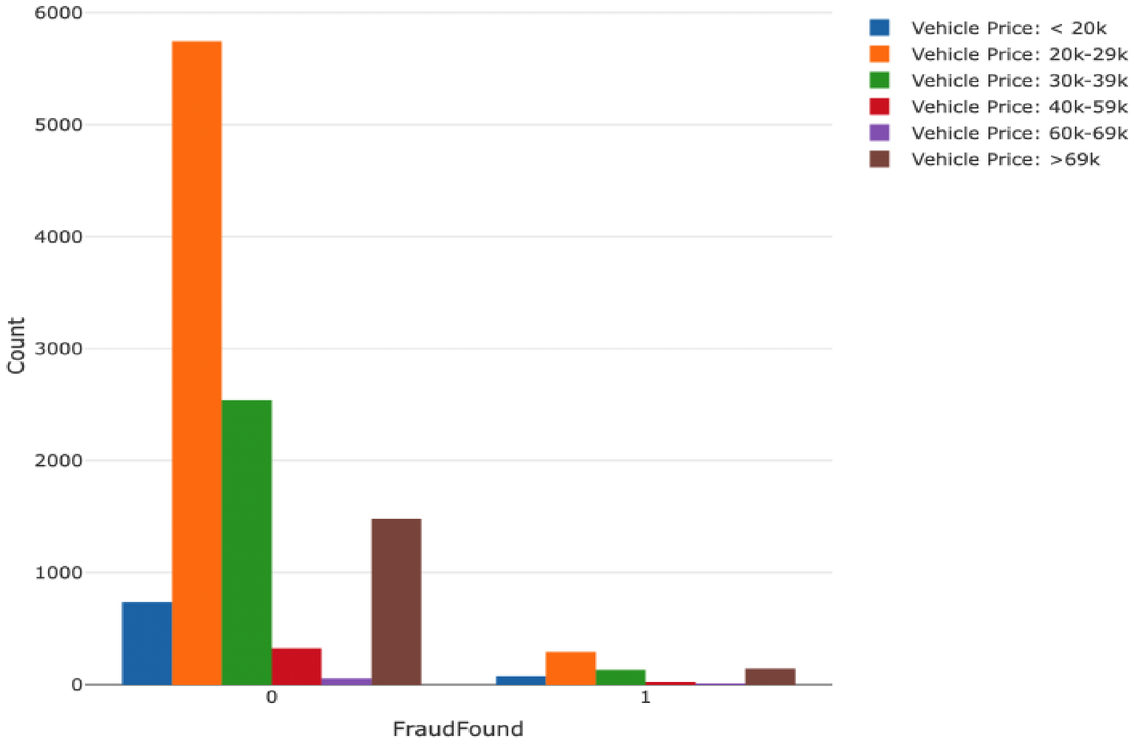
Institute and Faculty of Actuaries

The Data



8000
7000
6000
5000
4000
3000
2000
1000
0
Count
0
1
FraudFound
Fault: Policy Holder
Fault: Third Party

The Data



Institute and Faculty of Actuaries

Models

Institute
and Faculty
of Actuaries

Models

- Data split 75-25 for training and validation
- C5.0 trained using standard algorithm
- CART pruned using cost-complexity
- GBM, Random Forest and Neural Networks tuned using Cartesian Hyperparameter Grid Search

Grid Search Example – H2O

```
H2O Grid Details
================

Grid ID: gbm_grid
Used hyper parameters:
  -  col_sample_rate
  -  learn_rate
  -  max_depth
  -  ntrees
  -  sample_rate
Number of models: 144
Number of failed models: 0

Hyper-Parameter Search Summary: ordered by decreasing f1
  col_sample_rate learn_rate max_depth ntrees sample_rate         model_ids                 f1
1             1.0        0.1        25   5000         0.8 gbm_grid_model_66 0.26259541984732826
2             1.0        0.1        25   2000         0.8 gbm_grid_model_54 0.26259541984732826
3             1.0        0.1        25   8000         0.8 gbm_grid_model_78 0.26259541984732826
4             1.0        0.1        25  10000         0.8 gbm_grid_model_90 0.26259541984732826
5             0.8        0.1        25  10000         0.8 gbm_grid_model_89  0.2612085769980507

---
    col_sample_rate learn_rate max_depth ntrees sample_rate         model_ids                 f1
139             0.8        0.1        60   5000         0.6 gbm_grid_model_23 0.23591549295774647
140             0.8        0.1        60   2000         0.6 gbm_grid_model_11 0.23591549295774647
141             1.0        0.1        10   5000         0.6 gbm_grid_model_15   0.233983286908078
142             1.0        0.1        10   2000         0.6  gbm_grid_model_3   0.233983286908078
143             1.0        0.1        10   8000         0.6 gbm_grid_model_27   0.233983286908078
144             1.0        0.1        10  10000         0.6 gbm_grid_model_39   0.233983286908078
```

Institute and Faculty of Actuaries

Model Performance

- Evaluated using the following criteria
  - Accuracy
  - AUC
  - F1 Score
- All metrics based on **Confusion Matrix**
- AUC also related to **Receiver Operating Characteristics (ROC) Curve**

Institute and Faculty of Actuaries

Model Performance

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

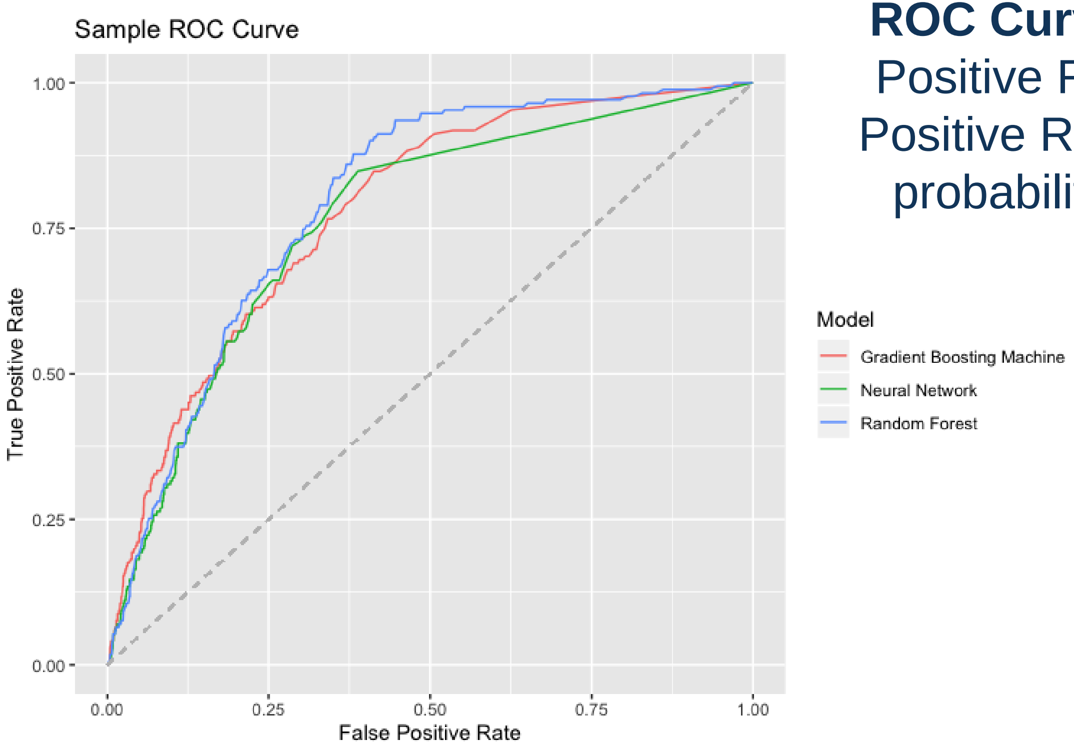
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1} = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

Institute and Faculty of Actuaries

Model Performance



ROC Curve: Plots True Positive Rate vs. False Positive Rate at different probability thresholds

AUC:
Area under ROC Curve

Measure of how well can a model distinguish between 2 classes

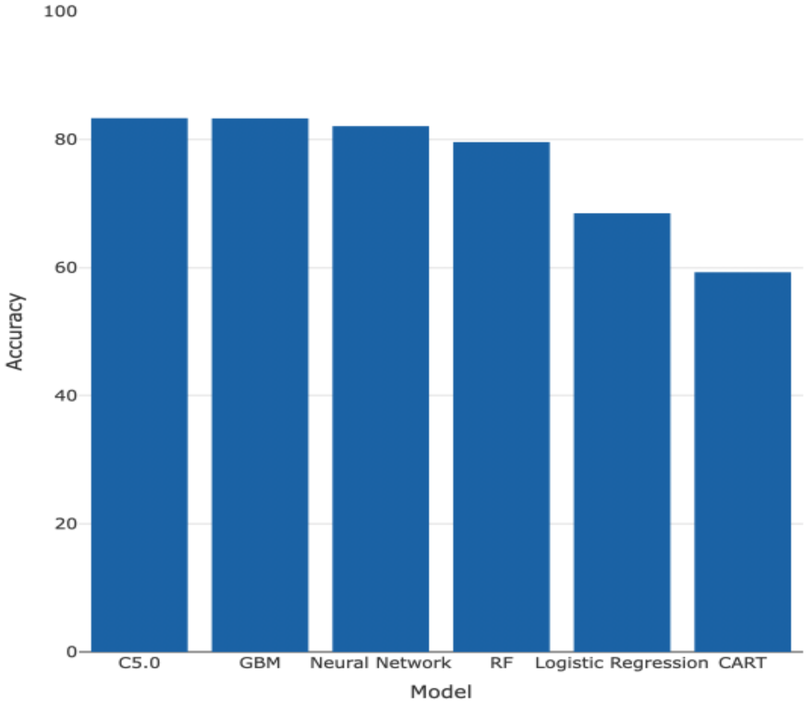
Institute and Faculty of Actuaries

Model Performance

- Evaluated using the following criteria
  - Accuracy
  - AUC
  - F1 Score
- All metrics based on **Confusion Matrix**
- AUC also related to **Receiver Operating Characteristics (ROC) Curve**

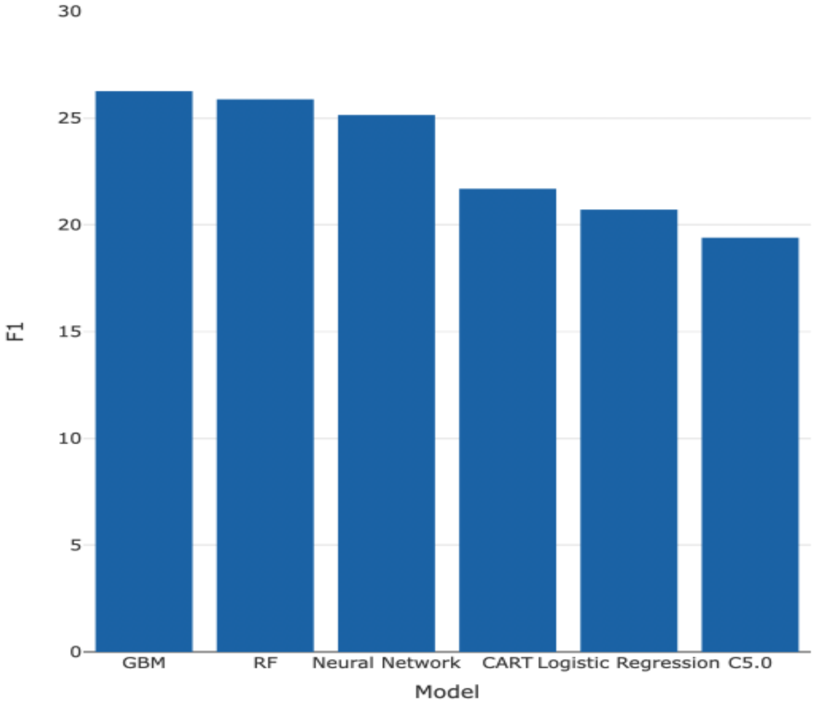
Institute and Faculty of Actuaries

Model Performance



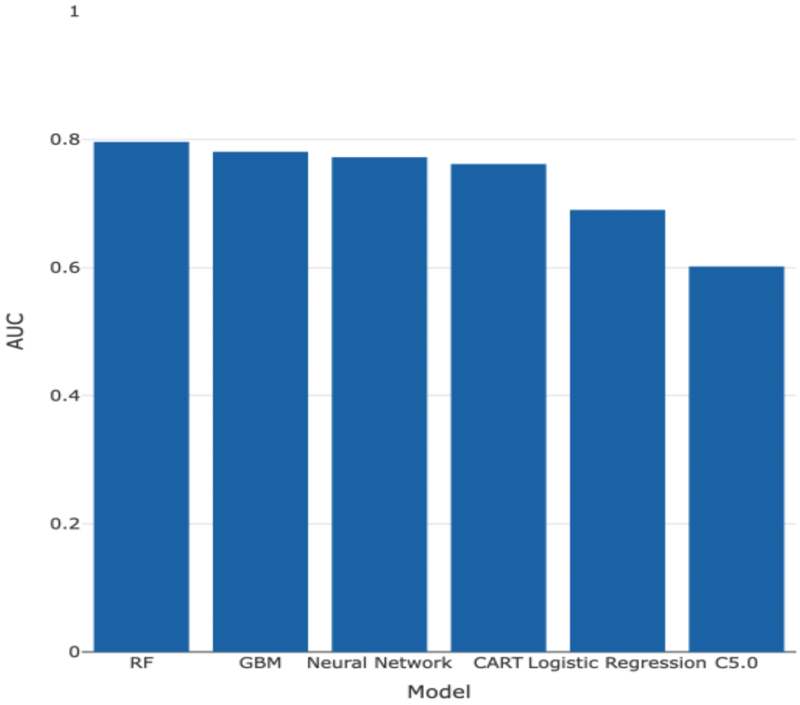
Institute and Faculty of Actuaries

Model Performance



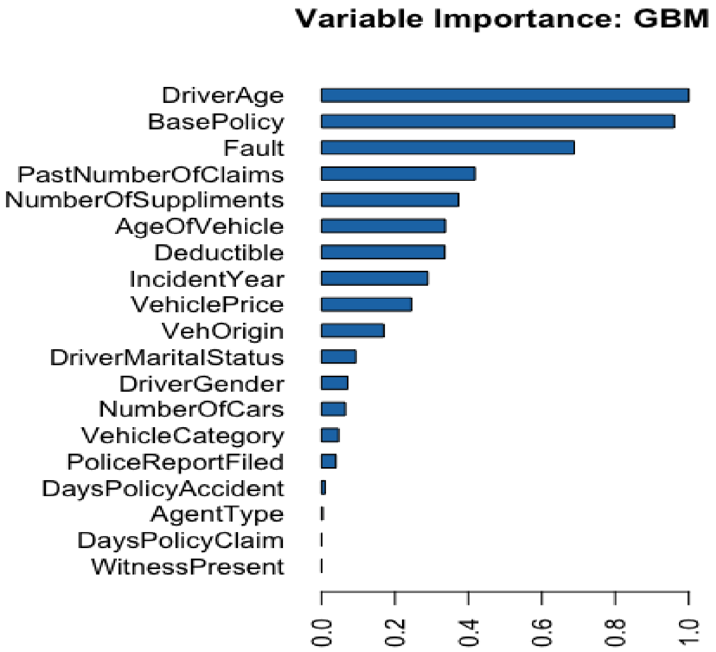
Institute and Faculty of Actuaries

Model Performance



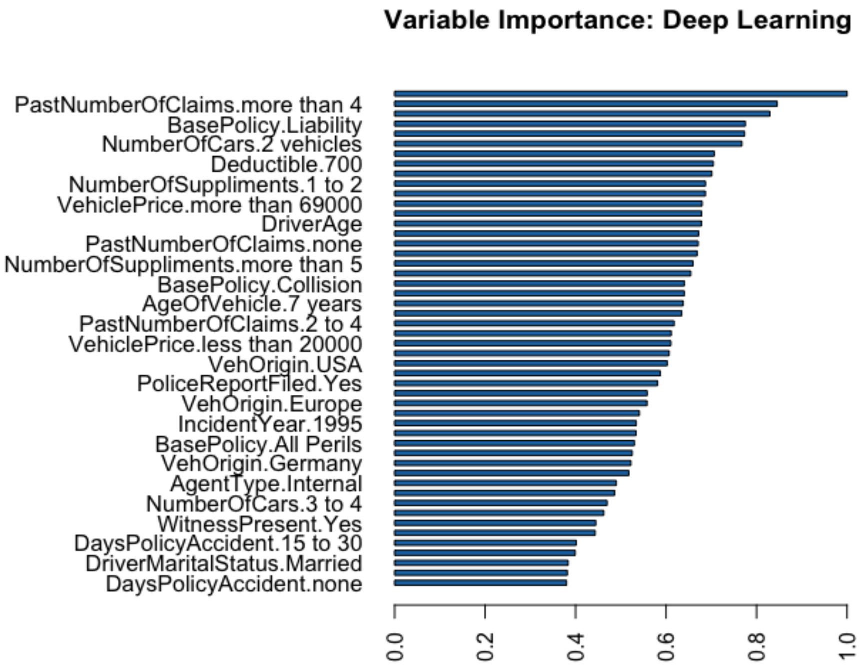
Institute and Faculty of Actuaries

Variable Importance



Institute and Faculty of Actuaries

Variable Importance



Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Key Takeaways & Conclusions

02 September 2019

Conclusions

- ML can be a powerful tool
- Results from classification models could be used to proactively flag claims as fraudulent and minimize unnecessary losses
- Models can also help understand customer behavior, eg. which groups contribute most to insurance fraud

Institute and Faculty of Actuaries

Questions

Comments

The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the author.

Institute and Faculty of Actuaries