Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modelling Applications

Authors: Galindo and Tamayo, 2000

Toby Riley | Connor McDowall | Ben Crosland

Agenda

Introduction (Toby Riley)

- Background
- Motivating question
- Research Question
- Literature Review
- Contribution

Data and Methodology (Connor McDowall)

- Introduction to modelling
- Introduction to algorithms
- Inductive principles
- Model construction, learning curves and complexity

Results (Ben Crosland)

- Model of results
- Comparison of results
- Conclusion, strengths and weaknesses

Motivating Question?

What are some core risks faced by banks and what does a financial crisis look like?

Credit risk

- Risk that the lender will not get paid all principal and interest on time as scheduled

Default risk

- Component of credit risk
- The risk that a borrower will not follow the agreed loan terms. This includes stopping payment, paying less or failing to maintain agreed financial covenants.

Attributes of a financial crisis

- Event triggers crisis
- Cash flow problems
- Loss of business confidence
- May be isolated by geography or global
- Asset prices fall
- Credit Crunch
- Increased risk

Background and Motivation

Several recent financial crisis highlight the importance of credit risk

1980’s

- Black Monday
- US savings and loans crisis
- Norwegian and Scandinavian banking crisis

1990’s

- Black Wednesday
- Mexico economic crisis
- Asian financial crisis
- Russian financial crisis

2000’s

- 2007 Financial Crisis
- COVID-19 pandemic 2020

Risk Exposure Measures

Early warning systems and risk decomposition and aggregation are the predominant ways to measure a financial institutions credit risk

Early Warning Systems

- Uses accounting data from financial statements to predict failure or problematic conditions

Financial EWS is a monitoring and reporting system that alerts for the probability of problems, risks and opportunities before they affect the financial statements of firms. EWSs are used for detecting financial performance, financial risk and potential bankruptcies. EWSs give a chance to management to take advantage of opportunities to avoid or mitigate potential problems.

Risk Decomposition and Aggregation

- Decomposes a financial institutions assets and liabilities into its exposure to pre-determined risk factors

Risk decomposition and aggregation has its roots in CAPM. It is a more effective measure of risk exposure as it allows for the source of the risk to be identified but having the correct pre-determined risk factors and accurately measuring them make it much more complicated.

Machine Learning Overview

Machine learning

- Machine Learning is the science (and art) of programming computers so they can learn from data
- Examples of ML that we see in our day-to-day life include the YouTube algorithm, Netflix recommendations, spam folder in your email.
- Examples in the finance industry include algorithmic trading, fraud detection and loan underwriting

Major brands such as YouTube and Netflix have adopted machine learning to understand consumer behaviours and recommend movies / videos

Freddie Mac

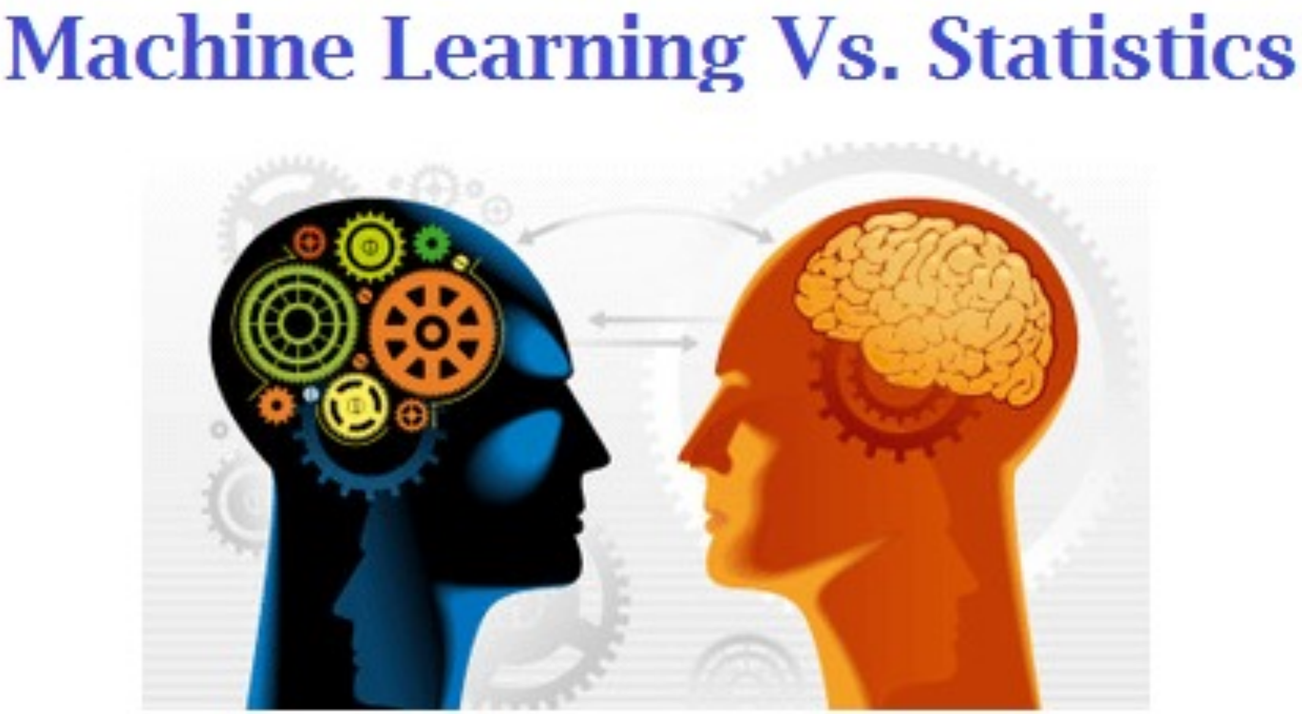
Fannie Mae

Freddie Mac and Fannie Mae both maintain and market large automated underwriting systems

Discussion Question:

Why has machine learning struggled so much to break into mainstream finance literature?

What is the difference between the two?



Research Purpose:

Making a comparative analysis on a variety of statistical and machine learning techniques to understand their strengths and weaknesses

Literature Review & Contribution

There has been a limited number of articles comparing different machine learning techniques ability to predict credit risk

1. **Using neural networks to predict bank failure**
   K. Tam, M. Kiang 1992
2. **Predicting bond ratings using neural networks**
   J. Maher, Tarun K. Sen 1997
3. **ML in financial crisis prediction**
   Wei-Yang Lin, Ya-Han Hu, Chih-Fong Tsai, 2012
4. **Comparison of techniques for credit scoring**
   Soner Akkoç 2012
5. **Use convolutional neural networks to predict mortgage defaults**
   Kvamme et al., 2018
6. **Empirical Asset Pricing via Machine Learning**
   Shihao Gu, Bryan Kelly, Dacheng Xiu, 2020

Contribution

1. Achieve an average error rate of 8.31% on a sample of 2000 with the optimal model
2. Multi-variety approach extends beyond prior literature
3. Introduce the modelling methodology of studying error curves
4. Help pave the way for ML to make it into mainstream finance

Data & Methodology

Connor McDowall

Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Introduction to Modelling: Overview

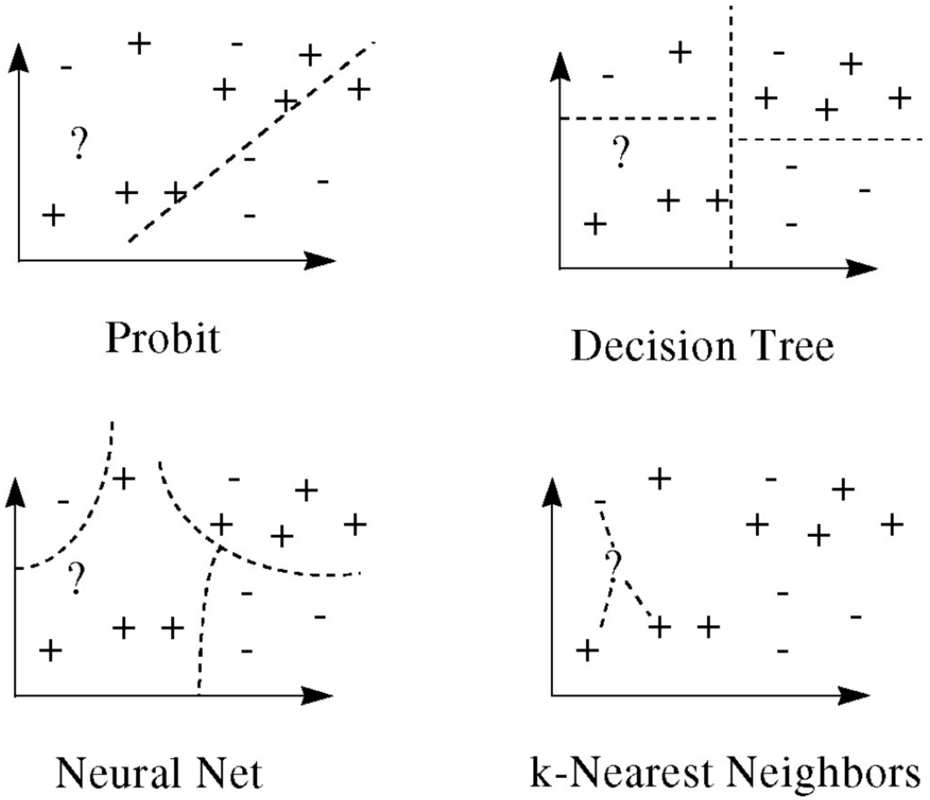
- The representation of patterns, regularities, or trends in financial or business data from statistical/mathematical models persists in finance
- Interdisciplinary techniques combining statistics and machine learning algorithms required to inform analysis on complex real-world data:
  - Noise
  - Non-linearity
  - Idiosyncrasies
- Increased relevance from:
  1) Corporate & Government Financial Databases
  2) Mature Statistical and Machine Learning Technologies
  3) Affordable Computing Resources

Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Introduction to Algorithms: Overview

- Algorithm selection relies on the nature and characteristics of the dataset e.g., Intrinsic noise, complexity, relationship-type.
- Significant variation in basic structure, parameters, and optimisation landscapes between algorithms:
    - **Traditional statistics: Probit Linear Regression (Top Left)**
    - **Modern statistics: k-Nearest Neighbors (Bottom Right)**
    - **Decision trees and rule-based induction methods: Decision Tree (Top Right)**
    - **Neural networks and related machines: Feed Forward Neural Network (Bottom Left)**
    - Bayesian Inference and Networks
    - Model combination methods: boosting and bagging
    - Genetic algorithms and intelligent agents
    - Fuzzy logic, fractal sampling, and hybrid approaches



Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Introduction to Algorithms: CART

- Classification and Regression Trees (CART) make decisions based on a branching structure
- Binary predictions start at the root node
- Each node has three attributes:
  - Samples: Count on number of sample node applies e.g., 2000 mortgage loans
  - Value: Number of instances per class applies on the node: 1500 default, 500 No default
  - Gini/Entropy: Impurity measure with purity (Gini/Entropy = 0) representing all training instances it applies to belong to the same class
- Partitions subsets by feature (k) and threshold (t_k), searching to produce the purest subsets
- Recursive, producing new leaf nodes until reaches maximum depth/allowed/purity
- Density in this instance is a toolset feature controlling

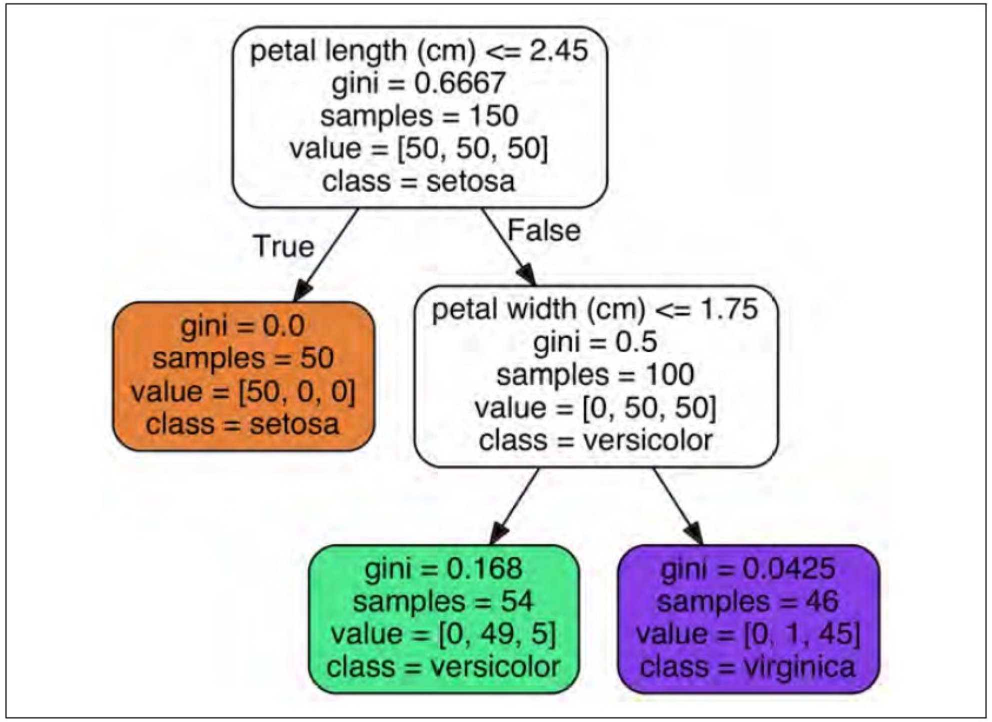


Figure 6-1. Iris Decision Tree

Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Introduction to Algorithms: Neural Net

- A feedforward artificial neural network (ANN) is a series of layered perceptrons
- A linear threshold unit (LTU) feeds a weighted sum of input values into a step/activation function to determine the output. A perception is a single layer of interconnected LTUs
  - Activation functions: sigmoid, hypertangent, and linear
- Perceptions utilize a training algorithm to assess the strength of connections between perceptions
  - Back propagation, steepest descent, conjugate gradient, modified newton, and genetic algorithm etc.
- A perception makes predictions on an instance one at a time, re-enforcing the connection weights from incorrect LTU prediction to improve performance

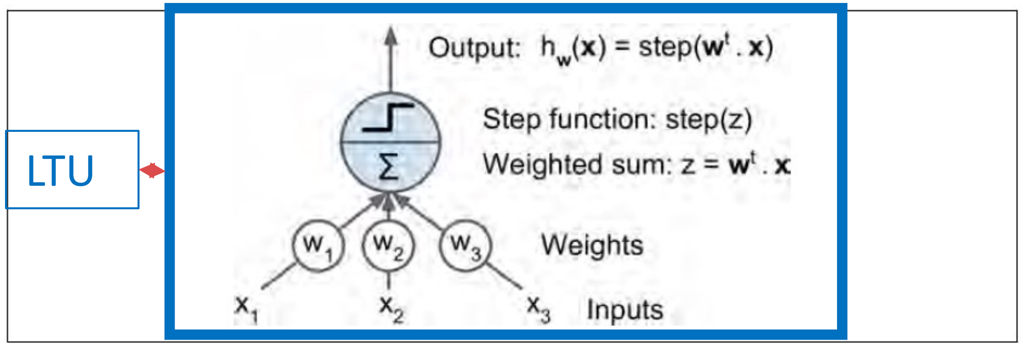


Figure 10-4. Linear threshold unit

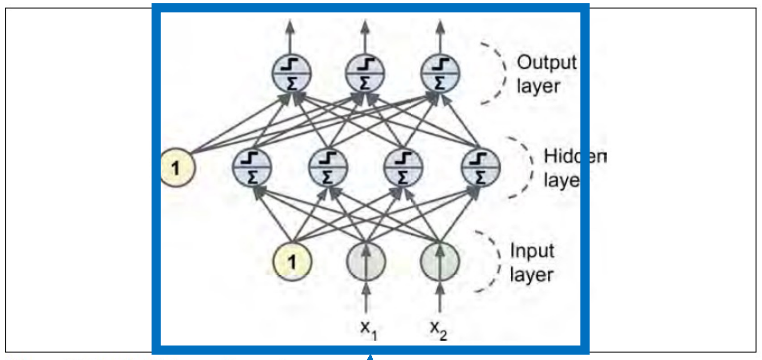


Figure 10-7. Multi-Layer Perceptron

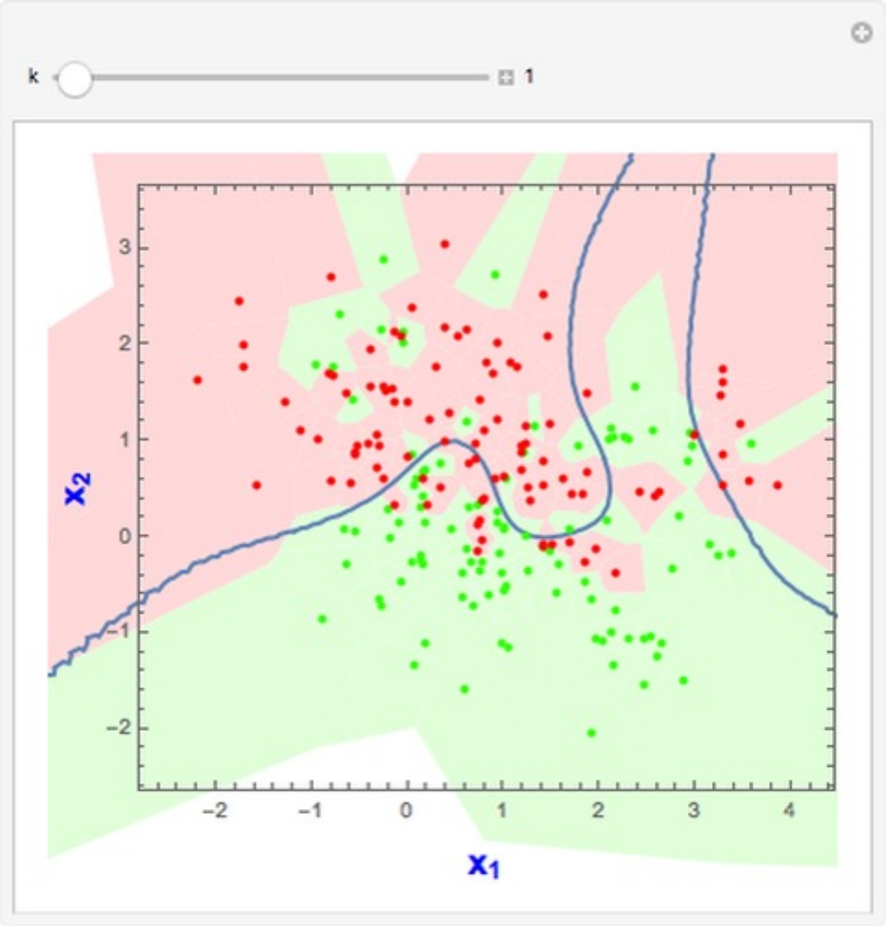
Multi-layer Perception

Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Introduction to Algorithms: k-Nearest Neighbor

- A simple, supervised machine learning algorithm which assumes similar instances exist in proximity
- The algorithm:
  1) Initializes K to the chosen number of neighbors
  2) Calculates the Euclidean distance to the nearest neighbors
  3) Performs a weighted average or majority vote to classify the instance
- Functional in low dimensions as prone to the 'curse of dimensionality'
- Significant slowdown issue as the complexity increases



Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Inductive Principles

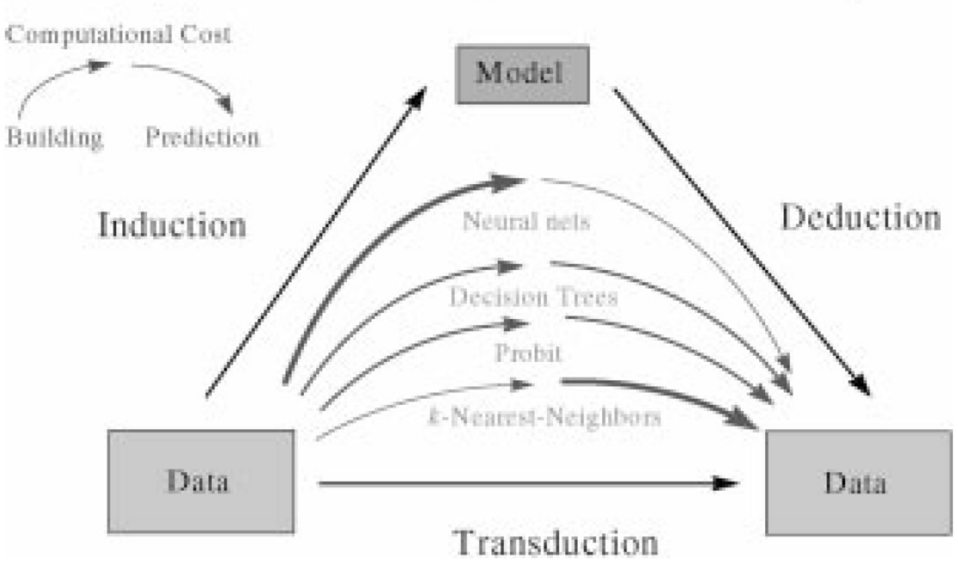
- Inductive principles guide the development of the most **parsimonious model/description** which captures the structural or functional relationship in the data: Keep all models or theories consistent with the data
- Traditional
  - **Maximum Likelihood Principle**
  - **Bayesian Inference:** Model selection by the maximization of posterior probabilities
  - **Minimization of Empirical Risk:** Accountability of model size and subsequent ability to generalize and finite sample behavior
- Modern (Occam's Razor): Most parsimonious model in fitting the data
  - **Minimum Description Length (MDL)**
  - **Komogorov Complexity:** Model with shortest/succinct computational representation

Strategy & Methodology

Multi-strategy Statistical Inference Approaches To Modelling

Inductive Principles

- Commonalities between traditional and modern approaches include:
  - Function approximation
  - Parameter/density estimation
  - Neural networks training methods
  - Data compression algorithms
- Model construction and computational cost closely intertwine (Right)
- The use of models in conjunction with interpretable theories/larger models requires:
  - Accuracy
  - Parsimonious (Succinct expression)
  - Non-trivial
  - Feasible
  - Transparent & interpretable
- Data quality & quantity dictate



Strategy & Methodology

Model Building & Analysis of Errors & Learning Curves

Model Construction

- Conceptually, statistical and machine learning models are not too different
- Computational and machine learning methods generalize parameter estimation underlying statistics using computational-based, and data-driven methods without relying on statistical assumptions
- Model construction and analysis relies on:
  1) Basic model parameter exploration: Preliminary,
  2) Analysis of importance/sensitivity: Not in this instance given number of variables in dataset
  3) Train/test/generalization and evaluation error analysis: Performance/confusion matrix for both testing and training (Right)
     - Data set divided into training (2000), testing (1000), and evaluation (1000) subsets. 22000 required

False positive

Table 1. Format of the performance matrix for a binary classification problem.

	Actual vs. predicted (performance matrix)			Total error
	Predicted (by model)			
	0	1	Total	
Actual 0	$x1$	y	$x1 + y$	Error for 0 $= y/(x1 + y)$
Actual 1	z	$x2$	$z + x2$	Error for 1 $= z/(z + x2)$
Total	$x1 + z$	$y + x2$	$x1 + x2 + y + z$	Global error $(z + y)/(x1 + x2 + y + z)$

False negative

Global Error

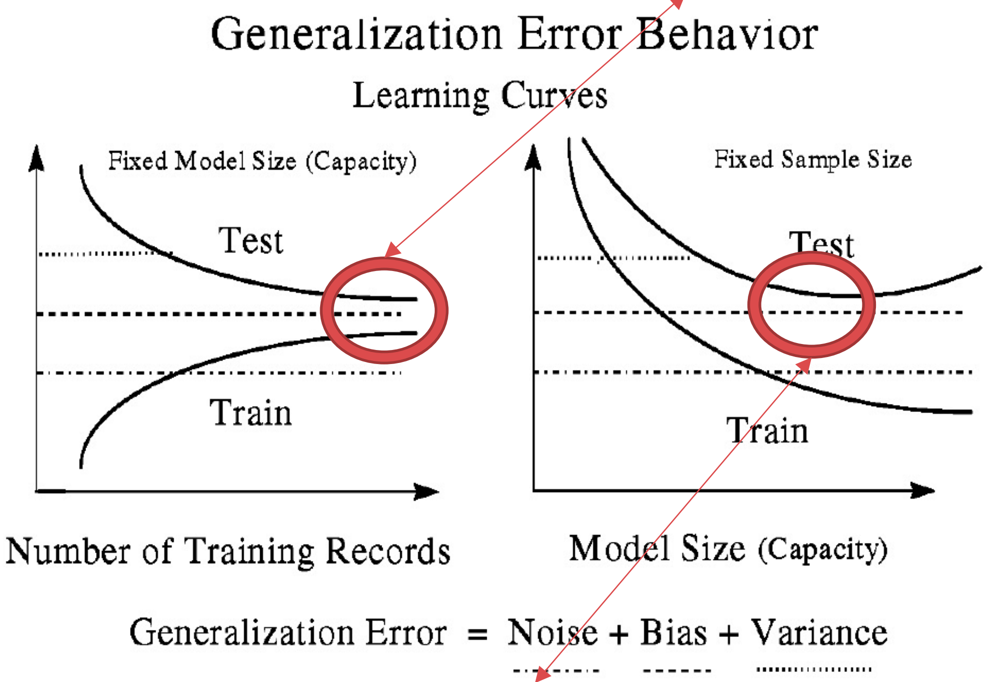
Strategy & Methodology

Model Building & Analysis of Errors & Learning Curves

Learning curves and complexity

- Model construction and analysis relies on:
  4) Analysis of learning curves and estimation of noise and complexity parameters
  - Computation of average values of training and testing (generalization) errors for giving values of model and sample size
  - Models the learning process and obtains rough estimates of complexity of and noise in the dataset after fitting simple algebraic scaling models
  - Process supports understanding of:
    1) Intrinsic complexity of the problem
    2) Quality of the data
    3) Insight into relationship between error rates, model capacity, and optimum training set sizes
    4) Useful for planning larger modelling relevant to production rather than exploratory data and view from **structural risk minimization** and **bias/variance decomposition** perspectives

Convergence to asymptotic determined by bias (model) and intrinsic noise of data



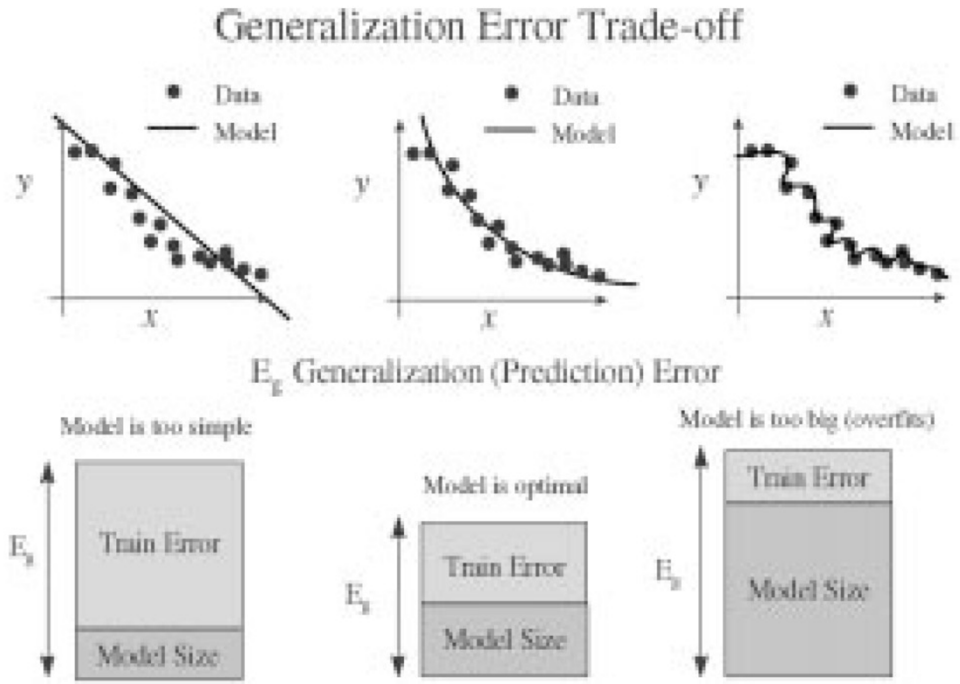
Optimal: Increase in size improves generalization error as less bias and fits data better until it overfitss

Strategy & Methodology

Model Building & Analysis of Errors & Learning Curves

Analysis of mortgage-loans learning curves

1) Build 30 models with different random samples from the original dataset for each dataset and model size
2) Average training and testing error rates
3) Fit errors to an inverse power law:
   - $E_{test} = \alpha + \beta/m^{\delta}$
   - α = noise/bias; β, δ = Complexity; m= sample size
   - Adapted/supported in prior research
   - $\delta \sim 1$ assumed as aligns with prior research on inverse power laws and theoretical models, provides reasonable fit
   - Caution given when using this inverse power law



Underfit (Left), Goldilocks (Center), Overfit (Right)

Strategy & Methodology

Application of the Analysis to a Financial Institution

Data Analysis, Preparation, Pre-processing

- Prediction of default in home mortgage loans
- Data provided by Mexico's security exchange and banking commission: Comision Nacional Bancaria y de Valores
- Possible set of 900,000 mortgage loans; 4000 mortgage loan from a single financial institution
- Average loan amount is 266,827 pesos ($33,300 U.S.) as of June 1996
- 24 attributes/variables (Credit_Amount, Overdue_Bal, Debt, Guarantee, DGuarantee1/2, Soc_Interest, Residential etc.)
- Target: Default (Binary; 0,1), considered only if no payments made in the last two months
- Default_Index: Condenses information about payment history and probability of payment

$$\boxed{P_{ij} = P(state_t = j \mid state_{t-1} = i).} \longleftrightarrow \boxed{\text{One-step probability}}$$

With the available information, the following one-step transition matrix P_{ij}^1 is calculated based on the frequency of each transition,

$$\boxed{P^1 = \begin{bmatrix} P_{00}^1 & P_{01}^1 \\ P_{10}^1 & P_{11}^1 \end{bmatrix}} \longleftrightarrow \boxed{\text{Transition matrix}}$$

This matrix is raised to power n (from 2 to 10) so that for every string of payment experiences the following variable is created,

$$\boxed{\text{Default\_Index} = \frac{\sum_{k=1}^{10} P_{ik1}}{10},} \longleftrightarrow \boxed{\text{Default\_Index}}$$

where P_{ik1} takes the value of P_{i1}^{11-k} if the account is in state i in the kth month.

Variables and Results

Ben Crosland

1 Probit Results

The probit regression estimates default risk with a 15.8% total error rate

Formula

$P\{Default = 1\} = \phi(\beta x_i)$

Cumulative normal distribution alternatives are used e.g. stepwise probit

βx_i
$= \beta_0 + \beta_1 Dguarantee1_i + \beta_2 Default\ index_i$
$+ \beta_3 Soc\ interest_i + \beta_4 Construction_i$
$+ \beta_5 Dguarantee1_i Default\ index_i$

Default threshold: 0.7

Where:

Dguarantee1 = Guarantee covers 100% of debt

Default index = variable reflecting the payment history and probability of payment

Soc Interest =

Construction = 1 for construction of a new house 0 if not

The probit model error is minimized at ~15% with a sample size of ~2000

Table III. Learning curve results for probit. The asymptotic value for the error rate is 15.02%. The standard deviation of the test error rate is in parenthesis

Model	Test error at $m = 2,000$	Noise/bias α	Complexity β	Optimum training sample size (recs)
Probit	15.13% (0.0047)	0.15025	1.80	1,804

Table IV. Actual vs. predicted results for Probit.

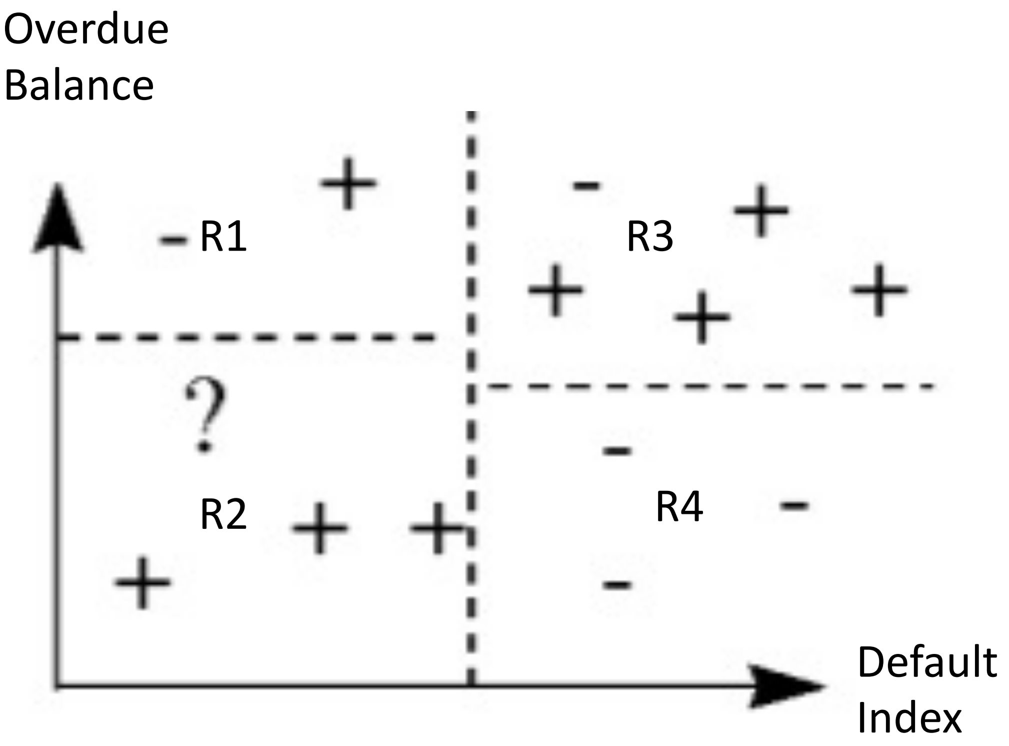
	Actual vs. predicted (performance matrix) Predicted 0	1	Total		
				Total error	15.80%
Actual 0	462	30	492	Error for 0	6.10%
Actual 1	128	380	508	Error for 1	25.20%
Total	590	410	1,000		

Error is significantly larger when estimating default

Sample Size	**50**	**150**	**280**	**700**	**1000**	**1500**	**2000**
Test Error	16.0%	15.8%	15.5	15.3	15.3	15.2	15.2

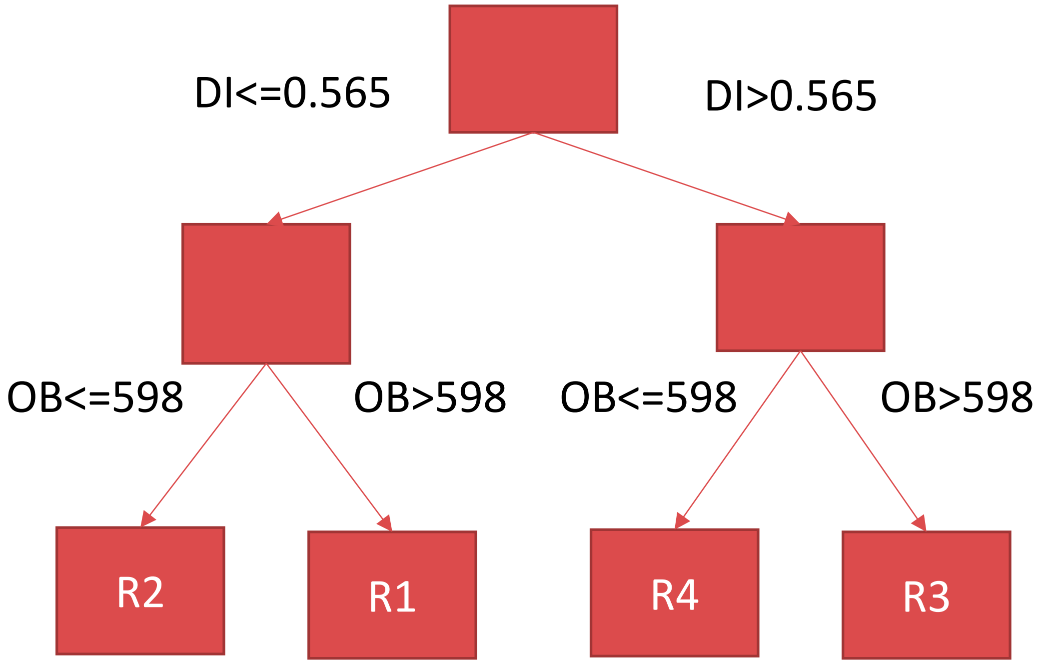
2 Decision-Tree Classification & Regression Tree (CART) Model

The Decision Tree CART model recap



Non parametric | Accurate | Interpretable

[TREE NODE 15 Records: Total 464, Target 471]
Default index <= 0.565 AND
Overdue Bal <= 598 AND
Debt > 21.275
THEN Default = 0 WITH misclassification error = 0.00632



2

Decision-Tree Classification & Regression Tree (CART) Model

The Decision Tree CART model had the lowest total error of 8.31%

- Reducing the density of the model results in a trade off between accuracy, model size and time
- The best tree is the tree with the emallest error in the test data set
  - This optimal subtree is obtained by a **pruning process**
  - Sub trees are generated by eliminating groups of branches based on a complexity / error trade off to **avoid over fitting**

Table V. Accuracy vs. time trade off for CART models

Density	Tree size (best tree)	Tree size (largest tree)	Test error (%)	Time (secs)
0.2	25	25	10.5	13
0.15	25	25	10.5	13
0.1	35	41	9.8	13
0.05	39	45	7.5	14
0.025	81	89	7	17
0.01	77	121	6.9	20
0.005	109	189	6.5	24
0	161	299	6.7	27

Table VI. **Learning curve results for different model sizes**

Size # of nodes	Test error at $m = 2,000$		Noise/bias α	Complexity β	Optimum training sample size (recs)
20	10.74%	(0.0055)	0.10400	6.36	6,357
40	9.13%	(0.0066)	0.08592	11.65	11,646
80	8.45%	(0.0060)	0.07591	18.13	18,127
100	8.41%	(0.0051)	0.07413	20.19	20,186
120	8.31%	(0.0058)	0.07312	21.68	21,675
200	8.31%	(0.0065)	0.07668	20.69	20,689
300	8.87%	(0.0075)	0.08230	17.16	17,160
400	8.97%	(0.0075)	0.08272	16.88	16,876

Table VII. Performance matrix for the tree with 120 nodes

	Actual vs. predicted (performance matrix)				
	Predicted				
	0	1	Total	Total error	9.10%
Actual 0	433	59	492	Error for 0	11.99%
Actual 1	32	476	508	Error for 1	6.29%
Total	465	535	1,000		

3 Neural Net Results

A preliminary analysis was taken place to identify the best neural net combinations…

- Conducted a number of preliminary neural networks combinations with different activation functions, training algorithms, and number of iterations
- **Train error: 10.28% - 18.97%**
- **Test Error: 9.88% - 19.11%**

Model selected for second stage analysis

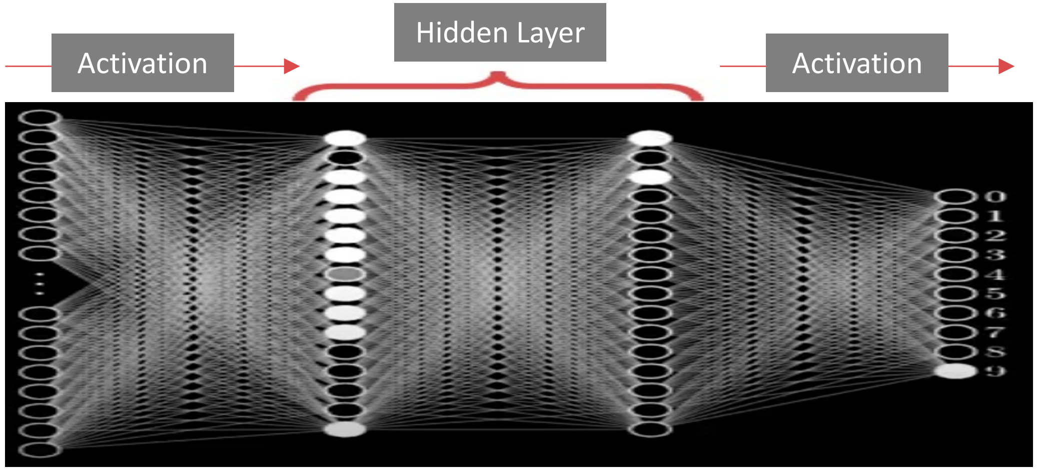


Table VIII. Preliminary exploration for neural networks

Number of nodes	Activation function	Training algorithm	Number of iterations	Train error	Test error
8	Sigmoid	Back propagation	900	18,97%	19,11%
8	Sigmoid	Steepest descent	96	11,72%	10,84%
8	Sigmoid	Conjugate gradient	46	10,39%	9,88%
8	Sigmoid	Modified newton	36	11,26%	10,44%
8	Sigmoid	Genetic algorithm	9	13,14%	12,15%
8	Linear	Back propagation	900	13,23%	12,68%
8	Linear	Steepest descent	27	12,05%	11,10%
8	Linear	Conjugate gradient	20	12,00%	11,11%
8	Linear	Modified newton	21	11,99%	11,12%
8	Linear	Genetic algorithm	9	15,48%	13,82%
8	Hypertangent	Back propagation	900	13,71%	13,38%
8	Hypertangent	Steepest descent	156	10,86%	10,36%
8	Hypertangent	Conjugate gradient	35	10,28%	10,07%
8	Hypertangent	Modified newton	41	10,43%	9,92%
8	Hypertangent	Genetic algorithm	9	13,84%	13,06%

3

Neural Net Results

… the neural net was refined by finding the optimum architecture and number of iterations…

How do different architectures (number of nodes) impact the test error?

Table IX. Results for neural nets of different sizes trained for a fixed number of batch iterations (25)

Size	Test error at $m = 2,000$		Noise/bias α	Complexity β	Optimum training sample size (recs)
2	11.00%	(0.0032)	0.10723	5.69	5,689
4	11.04%	(0.0033)	0.10776	5.23	5,233
6	11.09%	(0.0035)	0.10749	6.36	6,357
8	11.09%	(0.0033)	0.10768	6.92	6,916
16	11.15%	(0.0032)	0.10847	6.20	2,877

Hidden Layers

- **2 hidden layers minimizes the testing error**
- **Testing error increases presumably due to model capacity**

How does the number of training iterations impact the test error?

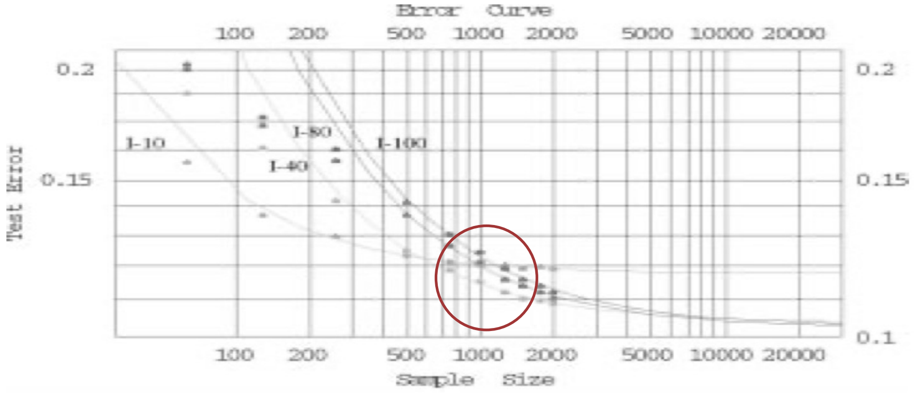


Figure 11. Generalization learning curves for 8-node neural nets with 10, 40, 80, and 100 iterations.

Table X. Neural nets with 8 nodes in hidden layer.

Iterations	Test error at $m = 2,000$		Noise/bias α	Complexity β	Optimum training sample size (recs)
10	11.92%	(0.0032)	0.11785	2.77	2,773
25	11.09%	(0.0033)	0.10768	6.92	6,916
40	10.89%	(0.0033)	0.10365	10.86	10,864
80	11.05%	(0.0034)	0.10240	17.57	17,567
100	11.19%	(0.0038)	0.10284	20.02	20,025

3 Neural Net Results

… to find the optimal neural net provided a total error of 15.4% and a 19.5% default

Table XII. Neural net with 16 nodes and 80 iterations.

	Actual vs. predicted (performance matrix)				
	Predicted				
	0	1	Total	Total error	15.40%
Actual 0	437	55	492	Error for 0	11.18%
Actual 1	99	409	508	Error for 1	19.49%
	536	464	1,000		

3 K-Nearest Neighbours

The K-NN estimates credit risk with a 17.7% total error rate

- Allows data itself to provide the ‘model
- Requires a large dataset
- **A practical problem in applying k-NN to the mortgage loan data set arises since the number of records is small**

Table XIII. Error rates for k-NN.

k	Test error for 750 records	
2	20.05%	(0.0077)
4	18.32%	(0.0062)
8	17.25%	(0.0053)
16	15.53%	(0.0098)
20	15.05%	(0.0059)
24	14.95%	(0.0049)
28	15.03%	(0.0050)
32	15.05%	(0.0050)

Table XIV. Performance matrix for $k = 24$.

	Actual vs. predicted matrix				
	Predicted				
	0	1	Total	Total error	17.70%
Actual 0	437	61	492	Error for 0	12.40%
Actual 1	116	392	508	Error for 1	22.83%
	547	453	1,000		

Comparison of models

Summary of best models performance, complexity and optimal sample size

Model	Test error (2,000 recs.)	Noise/bias α	Complexity β	Optimum training sample size (recs.)
CART (120 nodes)	8.3%	0.073	21.7	21,675
Neural net (16,80)	11.0%	0.102	18.1	18,165
k-NN	14.95% (1,000 recs.)	–	–	–
Probit	15.13%	0.150	1.80	1,804

- This model also has the smallest noise/bias parameter and, the highest complexity
- Confirms the hypothesis that the good models exploit the data more fully and converge more slowly to their asymptotic value
- Requires a sample size of ~22,000 to optimise the results for this financial institution

Table XVI.

Model (s)	Absolute (%)	Error for 0 (%)	Error for 1 (%)
CART	9.10	11.99	6.30
k-NN	17.70	12.40	22.83
NeuralNet	15.40	11.18	19.49
Probit	15.80	6.10	25.20
CART AND k-NN	14.10	3.66	24.21
CART AND Neural Net	12.60	3.86	21.06
CART AND Probit	14.80	3.25	25.98
k-NN AND Neural Net	17.30	7.11	27.17
k-NN AND Probit	16.50	3.86	28.74
Neural Net AND Probit	15.80	4.88	26.38
CART OR k-NN	12.70	20.73	4.92
CART OR Neural Net	11.90	19.31	4.72
CART OR Probit	10.10	14.84	5.51
k-NN OR Neural Net	15.80	16.46	15.16
k-NN OR Probit	17.00	14.63	19.29
Neural Net OR Probit	15.40	12.40	18.31
Mayority rule (CART, NN, k-NN)	13.20	9.35	16.93

Conclusion

- Analyses the performance of four different algorithms and 9000 models
- The optimal model in terms of the lowest average error rate was found to be the CART – decision tree model at a rate of 8.3%
- Introduce a modelling method based on studying error curves
- Provide a framework for how future machine learning models can be used for risk assessment of a credit portfolio

Strengths and Weaknesses

Strengths

1. High quality data directly from financial institutions
2. 9000 models and four different ML algorithms is a rigorous approach.
3. Provided a strong framework for assessing credit risk

Weaknesses

1. Limited number of data from the same financial institution
2. The data was from Mexican financial institutions which would make the results difficult to apply directly to a developed economy