Applying Data Mining Techniques to Credit Cards Dataset

Presentation by Alice, Joe and Yaksh

Data Source & Problem Definition

Source: The study took payment data in October, 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank. Among the total 25,000 observations, 5529 observations (22.12%) are the cardholders with default payment.

Problem definition:

- Investigate and understand customers' default payment behavior in Taiwan.
- Aims to compare the predictive accuracy of data mining methods in determining the probability of default among these customers.

Why this problem is important?

1. Risk Management (help business manage risks effectively)
2. Customer Protection
    + identifying potential financial distress among customers.
    + offer proactive support or guidance to customers who might be at risk of defaulting
    + promoting financial well-being
3. Business Performance (which rely on credit sales/payment plans-> predict -> optimize revenue streams + minimize losses from non-payment)
4. Policy Implications (credit regulations, risk assessment standards, and consumer protection measures -> inform policy decisions.)

Training & Testing Dataset

- Data was randomly divided into two groups, one for model training and the other to validate the model.
- Total records: 30000 records
- Testing: 9000 records
- Training: 21000 records
  + More than testing to train model→ better performance
  +

Target variable & Features

->Concept: A payment default usually happens after multiple payments on a loan or other debt are missed.

-> In study, Default payment (Yes = 1, No = 0) is response variable

-> In project, we choose Default payments is target attribute.

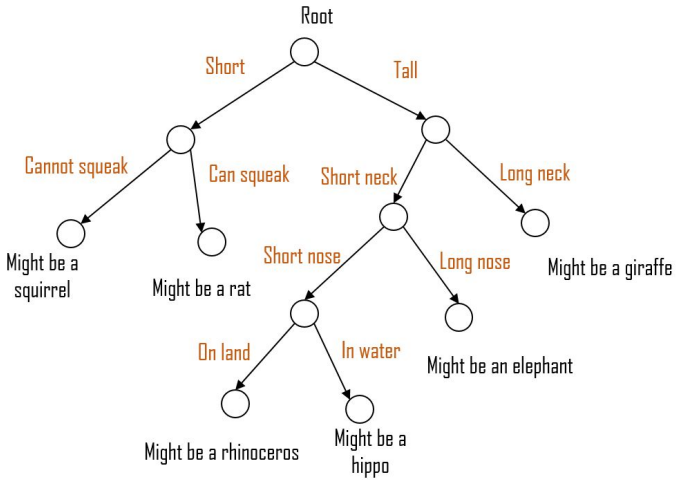
Features

23 features in the dataset.

- Amount of credit history
- Gender
- Education
- Marital Status
- Age
- History of Payment (April to September 2005)
- Amount of bill statement
- Amount of previous payment

Decision tree general concepts

A decision tree is a decision support hierarchical model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.



Decision tree Visualization

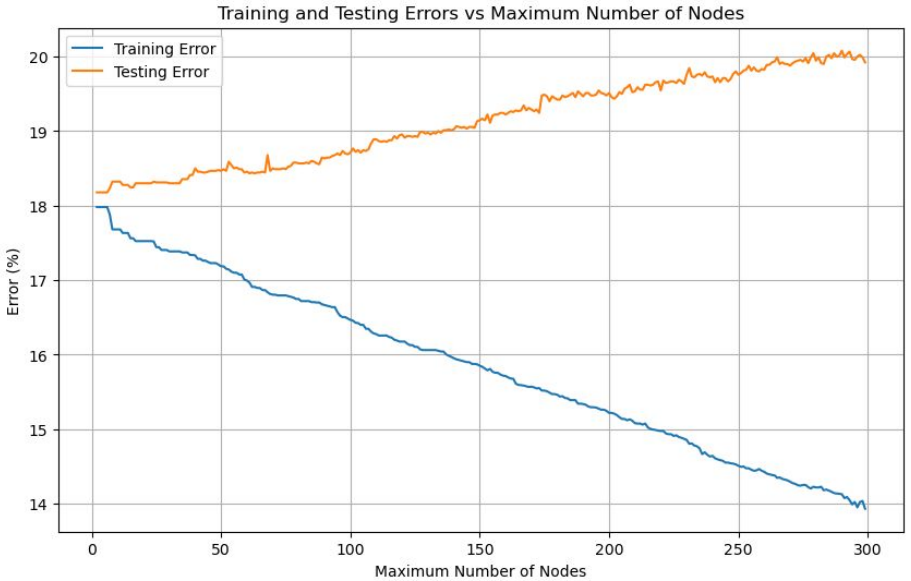
Accuracy: 71.6%
Testing Error: 28.4%

		Actual	
		0	1
Predicted	0	5664	1210
	1	1345	781

	precision	recall	f1-score	support
0	0.83	0.81	0.82	7009
1	0.37	0.40	0.38	1991
accuracy			0.72	9000
macro avg	0.60	0.60	0.60	9000
weighted avg	0.72	0.72	0.72	9000

Influence of varying nodes

The process of decision tree construction is to make the training error of the current training sample as small as possible, but this is also the main reason for overfitting.



Advantages of decision tree

1. Intuitive: The results of the decision tree can be interpreted intuitively and are easy to understand.

2. High computational efficiency: decision tree algorithm has relatively low computational complexity and does not require a lot of memory.

3. Feature selection: Decision tree algorithms can automatically select the most important features, which is very helpful for understanding and improving the predictive performance of the model.

Disadvantages of decision tree

1. Overfitting: Decision trees can be too complex, resulting in overfitting. This can degrade the performance of the model on unknown data.

2. Noise: Decision trees are very sensitive to noise, a small amount of noise can cause the performance of the decision tree to be greatly reduced.

3. Continuous attributes error: For continuous attributes, the decision tree needs to set thresholds for segmentation, which can lead to unstable results.

Support Vector Machine (SVM) General Concepts

- Hyperplane
- Support Vectors
- Margin
- Kernel
- Hard Margin
- Soft Margin
- C
- Hinge Loss

$$\text{Linear}: \ K(w,b) = w^T x + b$$

$$\text{Polynomial}: \ K(w,x) = (\gamma w^T x + b)^N$$

$$\text{Gaussian RBF}: \ K(w,x) = \exp(-\gamma ||x_i - x_j||^n$$

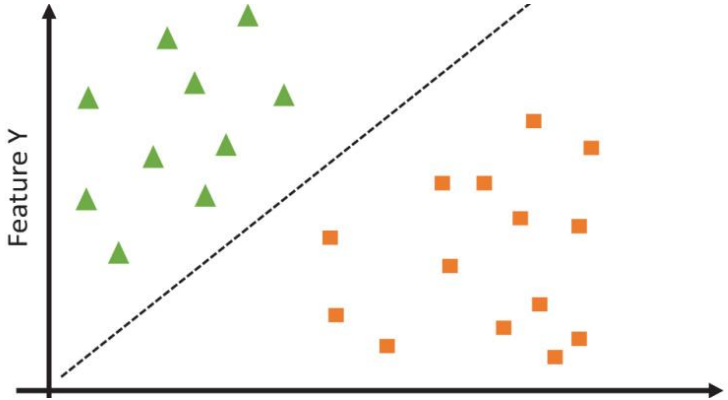
$$\text{Sigmoid}: K(x_i, x_j) = \tanh(\alpha x_i^T x_j + b)$$

Goal: Choose hyperplane to maximize distance of nearest data point on each side and penalize any misclassifications.

SVM Linear

Accuracy: 80.8 %
Testing Error: 19.2%

- Very suitable when data can be precisely linearly separated.



		Actual	
		0	1
Predicted	0	6797	1515
	1	212	476

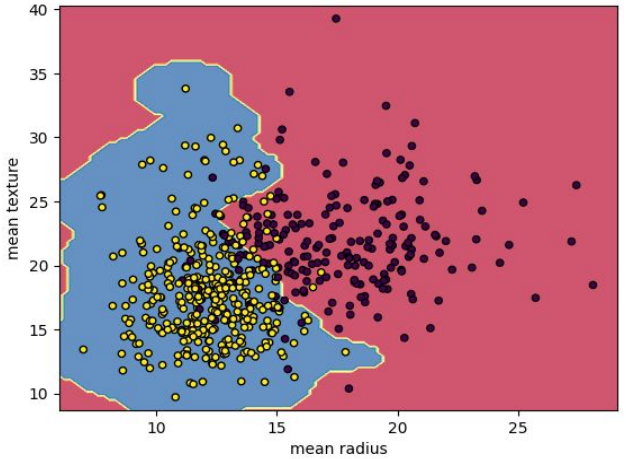
	precision	recall	f1-score	support
0	0.82	0.97	0.89	7009
1	0.69	0.24	0.36	1991
accuracy			0.81	9000
macro avg	0.75	0.60	0.62	9000
weighted avg	0.79	0.81	0.77	9000

SVM Radial Basis Function (RBF)

- Can model non-linear and complex relationships or clusters and it can create smooth and circular decision boundaries.

Accuracy: 81.4%

Testing Error: 18.6%



Breast Cancer Classifications with SVM RBF kernel

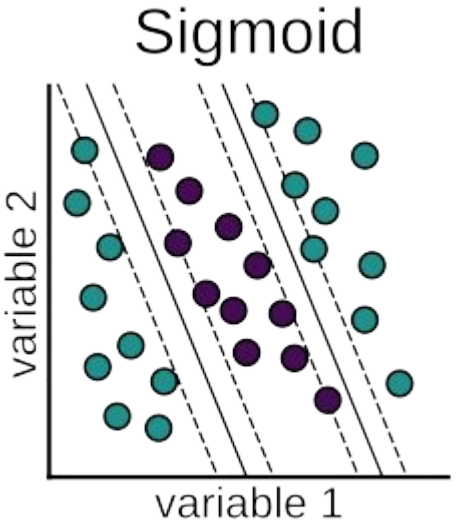
		Actual	
		0	1
Predicted	0	6676	1341
	1	333	650

	precision	recall	f1-score	support
0	0.83	0.95	0.89	7009
1	0.66	0.33	0.44	1991
accuracy			0.81	9000
macro avg	0.75	0.64	0.66	9000
weighted avg	0.79	0.81	0.79	9000

SVM Sigmoid

- Works well if data distribution looks like a logistic function.

Accuracy: 70.3%
Testing Error: 29.7%



		Actual	
		0	1
Predicted	0	5744	1406
	1	1265	585

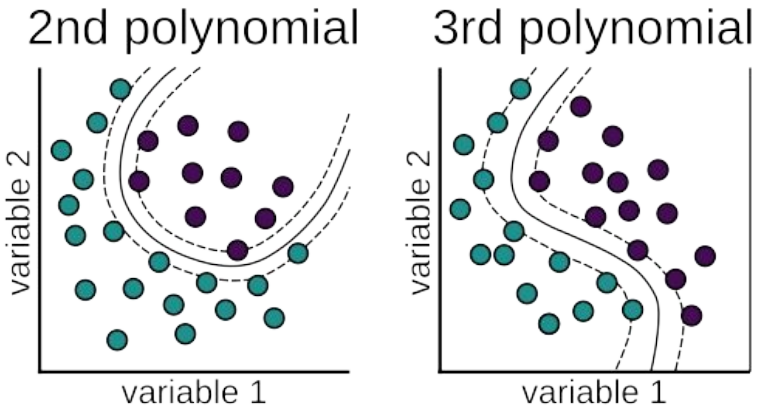
	precision	recall	f1-score	support
0	0.80	0.82	0.81	7009
1	0.32	0.29	0.30	1991
accuracy			0.70	9000
macro avg	0.56	0.56	0.56	9000
weighted avg	0.70	0.70	0.70	9000

SVM Polynomial (degree 4)

- Useful for non-linear patterns or interactions

Between features.

Accuracy: 79.7%
Testing Error: 20.3%



		Actual	
		0	1
Predicted	0	6764	1585
	1	245	406

	precision	recall	f1-score	support
0	0.81	0.97	0.88	7009
1	0.62	0.20	0.31	1991
accuracy			0.80	9000
macro avg	0.72	0.58	0.59	9000
weighted avg	0.77	0.80	0.75	9000

Advantages of SVM

- Effective in high dimensional cases.
- Robust to noise.
- SVM can handle irrelevant and redundant data better than many other techniques.
- Works well on complex small or medium sized datasets.

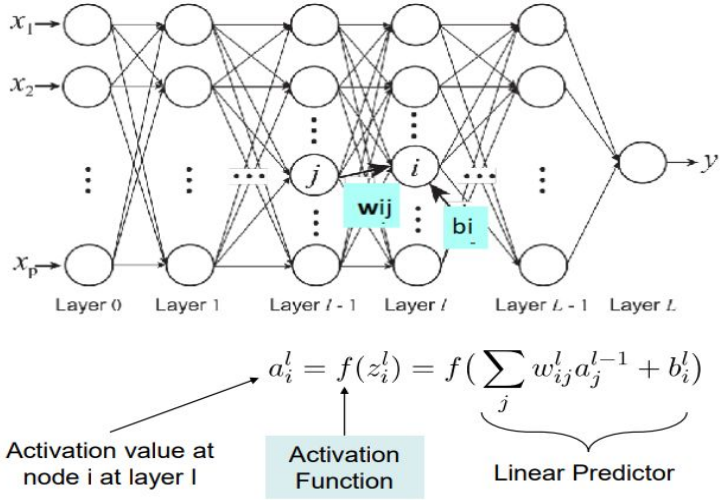
Disadvantages of SVM

- Computationally expensive during training.
- SVM is not suitable for large datasets.
- SVM does not perform well if there are overlapping cases.

Artificial Neural Network (ANN)

- Can be used to solve complex nonlinear problems.

Accuracy: 81.6%
Testing Error: 18.4%



		Actual	
		0	1
Predicted	0	6613	1264
	1	396	727

	precision	recall	f1-score	support
0	0.84	0.94	0.89	7009
1	0.65	0.37	0.47	1991
accuracy			0.82	9000
macro avg	0.74	0.65	0.68	9000
weighted avg	0.80	0.82	0.80	9000

Varying Number of Hidden Layers

ANN - Continued

During training, ANN did not converge. Weights of the neural network was not stable yet. Maximum iterations were 200.

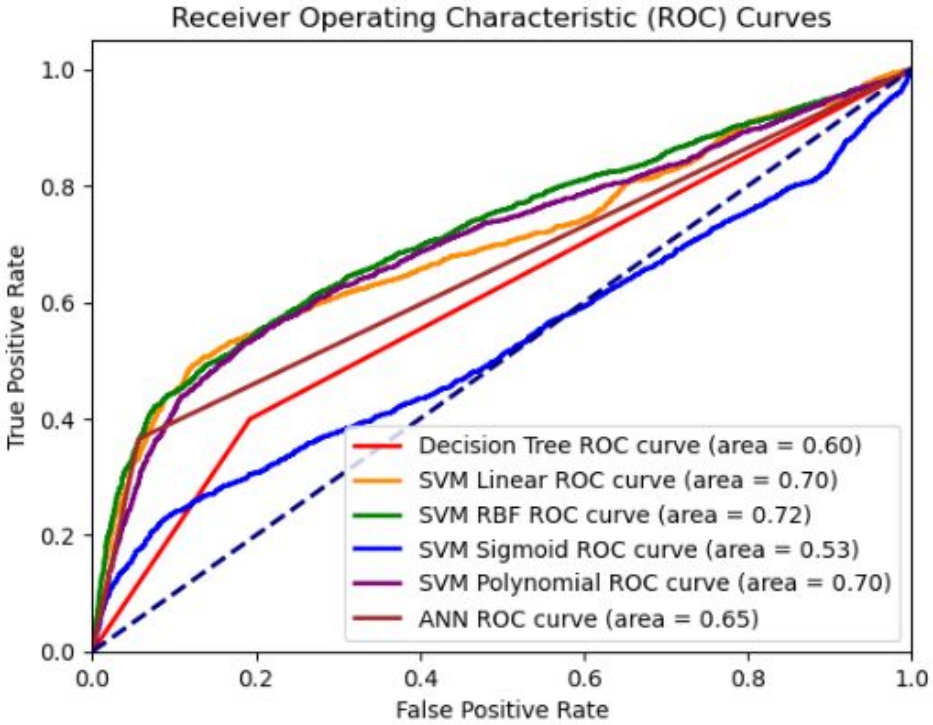
Disadvantages:

- Computationally expensive.
- Gradient descent may converge to a local minimum.
- Architecture depends on number of hidden layers, activation functions, loss function, optimization and how weights are updated.

Final Results

SVM RBF looks to be the best classifier throughout.

There are few points where ANN and SVM linear might slightly outperform SVM RBF.



Ethical Issues

Fairness and Bias: The model should not discriminate against individuals based on sensitive attributes such as race, gender, ethnicity, or socioeconomic status. It's essential to ensure that the model's predictions are fair and unbiased across different demographic groups.

Avoiding Reinforcement of Biases: Historical biases present in the data should not be perpetuated by the model.

Transparency and Explainability: Black-box models can lead to mistrust and raise concerns about fairness.

Ethical Issue - Continued

Data Privacy: Credit card payment data is sensitive personal information. Data anonymization and encryption techniques should be employed to protect individuals' privacy.

Model Accuracy and Reliability: False positives (incorrectly predicting someone will default) and false negatives (incorrectly predicting someone won't default) can have significant impacts on individuals' financial well-being.

Accountability and Oversight: There should be mechanisms in place to hold individuals and organizations accountable for the decisions made by the model.

What's next?

- Continue to investigate if ANN might be the best model. Or use different architectures such as RNN.
- Brainstorm on ways to reduce/eliminate bias.
- Establish causal relationships.
- Use better features (macro features such as economics) or (micro features such as lifestyle) and time series data of longer period.

Thank You

We are now open to questions.