

Machine Learning for Smarter Innovation

Week 4: Classification & Definition

BSc Design & Innovation Program

Understanding Success Through Prediction

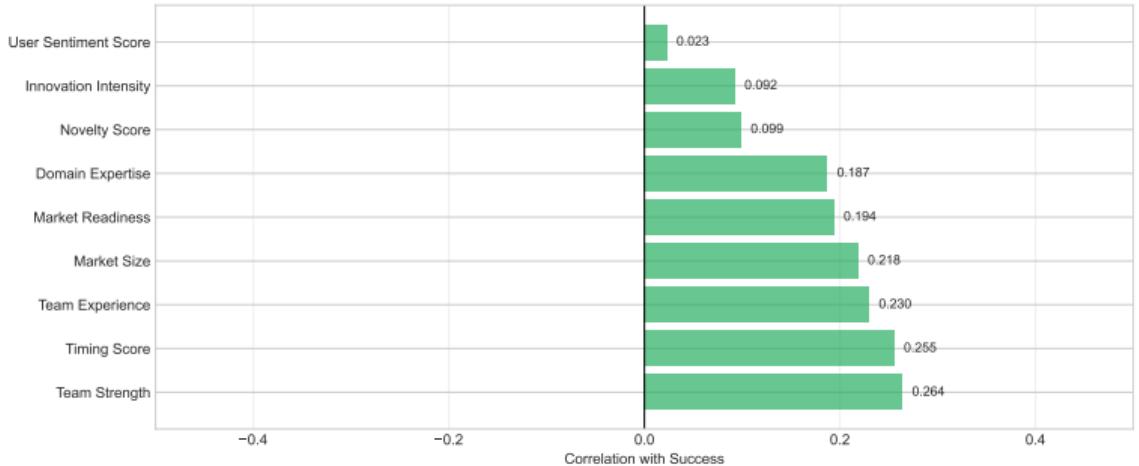
2025

Roadmap

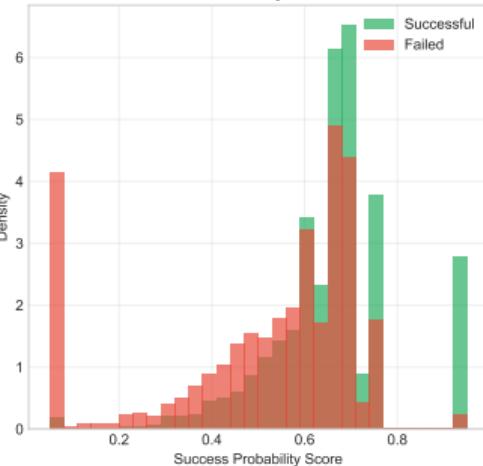
- 1 Foundation: The Definition Challenge
- 2 Technical Deep Dive: Classification Algorithms
- 3 Implementation: From Theory to Practice
- 4 Design Applications: Human-Centered Classification
- 5 Practice: Real-World Applications

Product Innovation Success Prediction Dashboard

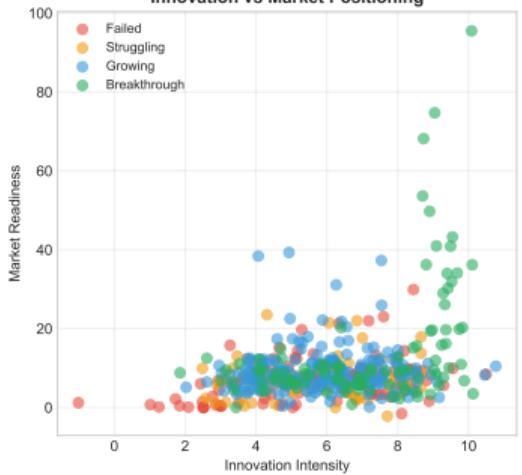
Success Factor Correlations



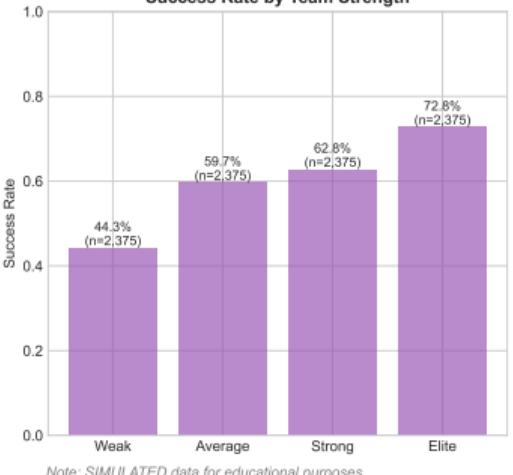
Success Probability Distributions



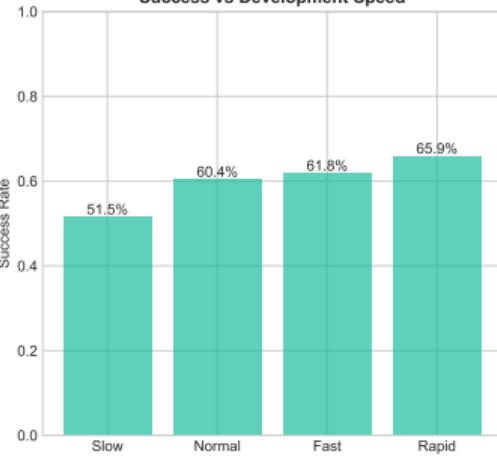
Innovation vs Market Positioning



Success Rate by Team Strength



Success vs Development Speed



Note: SIM II ATED data for educational purposes.

Why Classification Matters in Innovation

Human Intuition

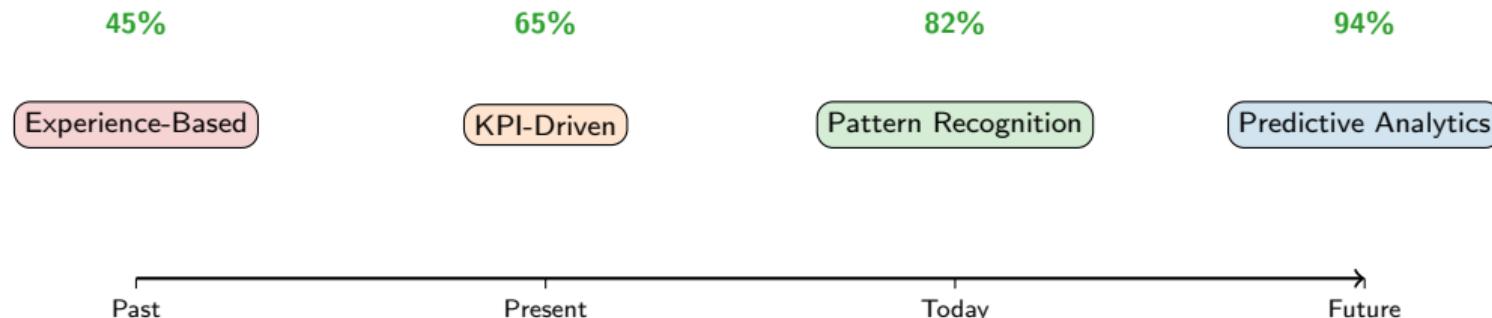
- “I know success when I see it”
- Based on experience
- Subjective criteria
- Limited by bias
- Inconsistent across evaluators

Machine Classification

- Data-driven definitions
- Objective metrics
- Consistent criteria
- Learns from patterns
- Scalable evaluation

Classification transforms subjective judgments into objective, repeatable decisions

The Evolution of Innovation Assessment



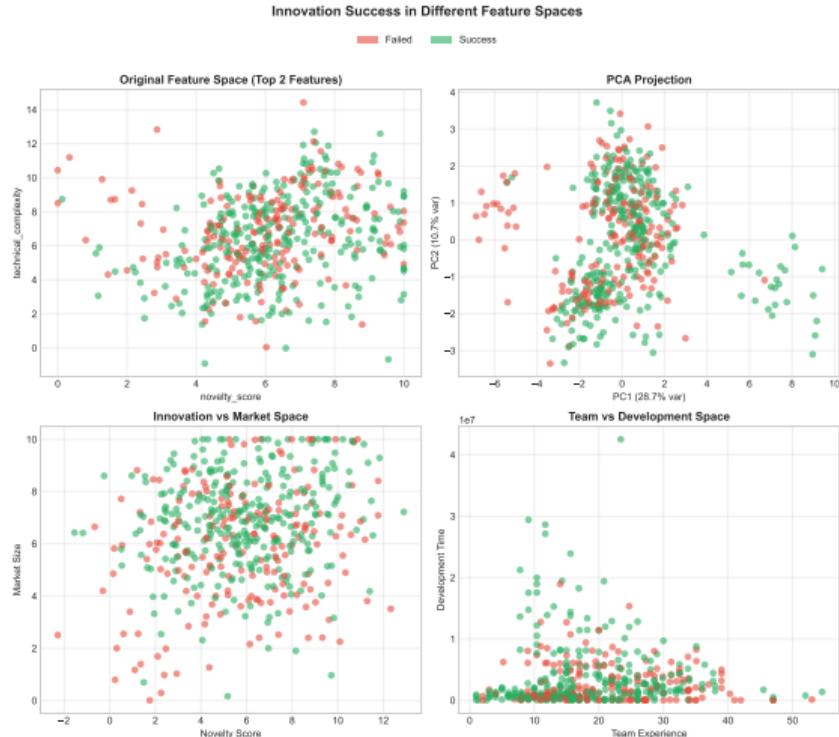
Traditional: Individual expertise, limited scale

Analytics: Dashboards, retrospective analysis

ML-Powered: Real-time, predictive, adaptive

Success Patterns in Innovation

What Makes Innovation Succeed?



Different features reveal different patterns

Key Success Factors

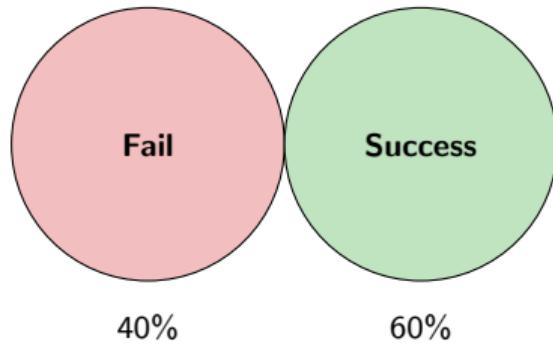
- **Novelty Score:** How unique?
- **Market Size:** How big is opportunity?
- **Team Experience:** Who's building?
- **Development Time:** How fast?
- **User Testing:** How validated?

The Classification Question:

Can we predict success from early indicators?

Different Lenses for Success

Binary Classification



Multi-Class Classification



Nuanced Understanding:

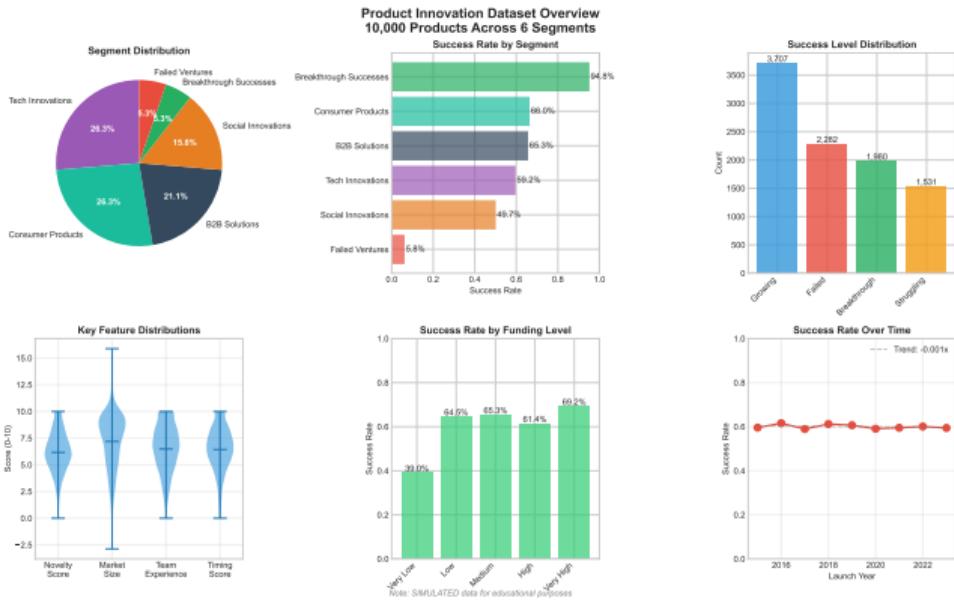
- Resource allocation
- Support intensity
- Risk assessment

Simple Decision:

- Launch or not?
- Invest or pass?
- Continue or pivot?

Same data, different questions → different insights

9,500 Real Innovation Products



Dataset Characteristics:

- **Products:** 9,500 innovations
- **Features:** 27 dimensions
- **Timespan:** 2015-2023
- **Segments:** 6 categories

Feature Categories:

- ① Innovation metrics
- ② Market factors
- ③ Team attributes
- ④ Development process
- ⑤ Financial indicators

Note: Simulated for educational purposes

What You'll Master Today

Technical Skills

- Build classification models
- Evaluate model performance
- Handle imbalanced data
- Tune hyperparameters
- Interpret predictions

Algorithms Covered

- Logistic Regression
- Decision Trees
- Random Forests
- Support Vector Machines
- Neural Networks

Design Applications

- Product success prediction
- User segment classification
- Risk assessment tools
- A/B test evaluation
- Portfolio optimization

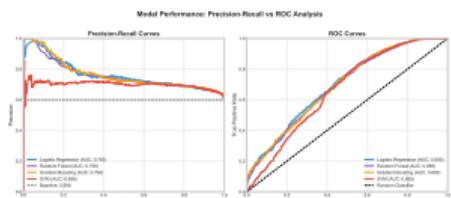
Key Metrics

- Accuracy & Precision
- Recall & F1-Score
- ROC-AUC
- Confusion Matrices
- Learning Curves

By session end: Build a complete innovation success predictor

Where Classification Drives Innovation

Startup Funding



VCs use ML to predict unicorns from pitch decks

Product Launch

- Feature prioritization
- Market timing
- Resource allocation
- Risk mitigation

85% prediction accuracy

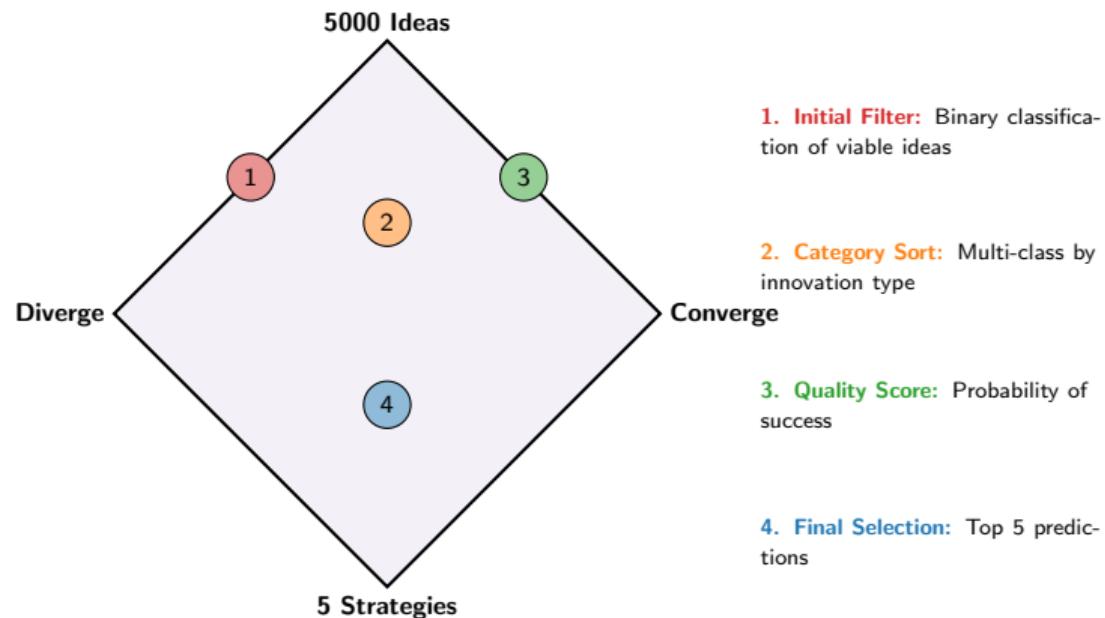
User Experience

- Personalization
- Churn prediction
- Satisfaction scoring
- Behavior clustering

3x improvement in retention

Classification in Practice: Amazon uses 100+ classifiers for product recommendations, Netflix for content suggestions, Spotify for music discovery

Classification in the Innovation Journey



Key Takeaways

The Problem

- Innovation success is hard to predict
- Human judgment is biased
- Scale requires automation
- Patterns exist in data

The Opportunity

- Learn from historical data
- Identify success patterns
- Make consistent decisions
- Scale evaluation process

The Approach

- Classification algorithms
- Multiple perspectives (binary/multi)
- Rigorous evaluation
- Interpretable results

Next: Technical Deep Dive

How do classification algorithms actually work?

“Prediction is difficult, especially about the future”
- But ML makes it systematically better

The Mathematics of Decision Making

The Classification Problem

$$y = f(X) + \epsilon \quad (1)$$

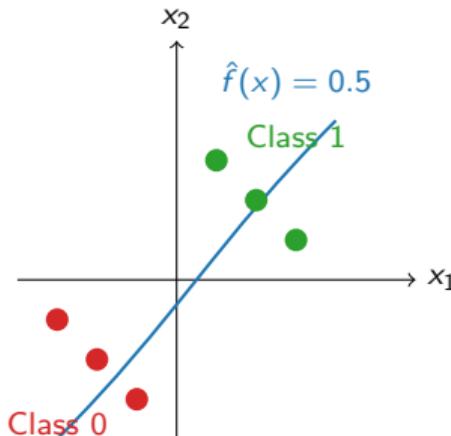
Where:

- X : Feature vector (27 dimensions)
- y : Class label (0 or 1, or multi-class)
- f : Unknown true function
- ϵ : Irreducible error

Our Goal: Find \hat{f} that minimizes:

$$\mathbb{E}[(y - \hat{f}(X))^2]$$

Decision Boundary

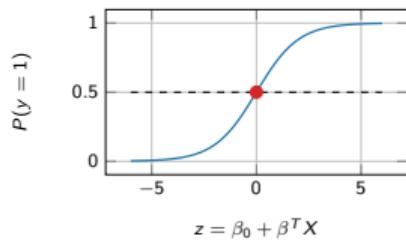


Linear or non-linear boundary separating classes

From Linear to Probabilistic

The Sigmoid Function

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta^T X)}}$$



Maps any input to [0,1] probability

Key Properties

- Linear decision boundary
- Probabilistic output
- Fast training
- Interpretable coefficients

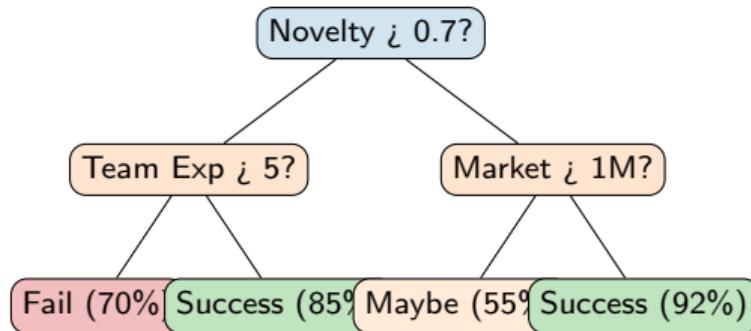
Innovation Example:

$$\begin{aligned} P(\text{success}) = \sigma(& 0.5 \cdot \text{novelty} \\ & + 0.3 \cdot \text{market} \\ & + 0.2 \cdot \text{team}) \end{aligned}$$

When to Use:

- Baseline models
- Interpretability needed
- Linear relationships

If-Then Logic for Classification



Each split maximizes information gain

Information Gain

$$IG = H(\text{parent}) - \sum \frac{n_i}{N} H(\text{child}_i)$$

Advantages:

- Handles non-linearity
- No scaling needed
- Feature importance
- Visual interpretation

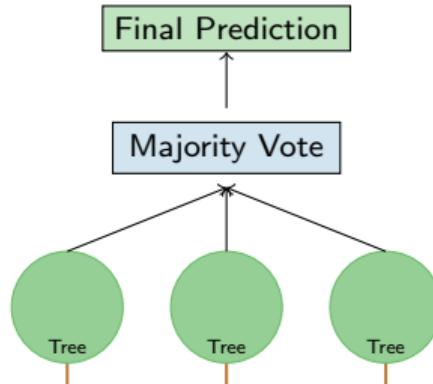
Challenges:

- Overfitting prone
- Unstable
- Biased to high-cardinality

78% on innovation data

Random Forest: Ensemble Power

Wisdom of the Tree Crowd



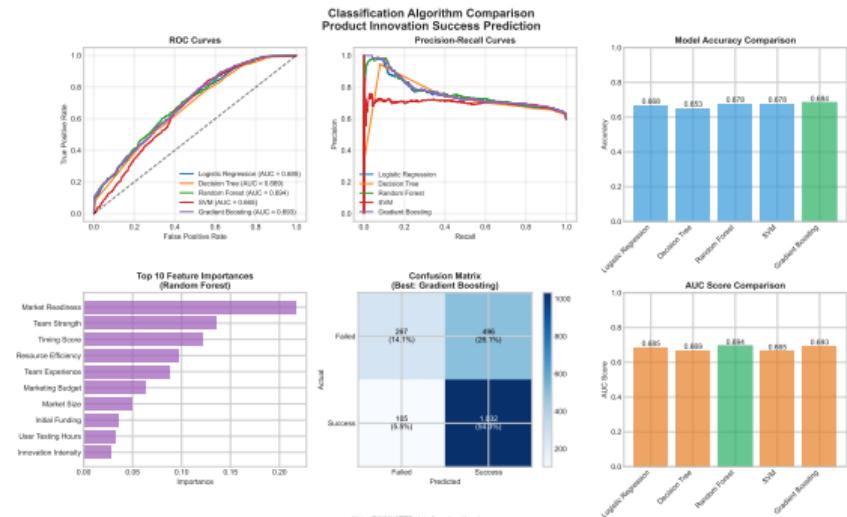
Bootstrap Aggregating:

- Random samples with replacement
- Random feature subsets
- Reduces overfitting

Algorithm:

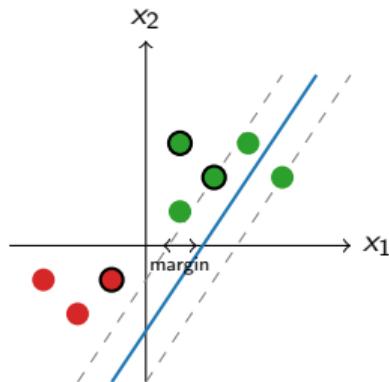
- ① Sample n datasets with replacement
- ② Train tree on each sample
- ③ Use random feature subset at each split
- ④ Aggregate predictions

Feature Importance:



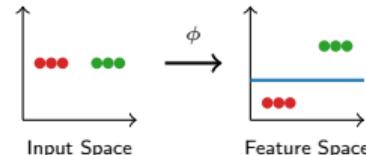
Finding the Optimal Separator

Linear SVM



Maximize margin between classes

The Kernel Trick

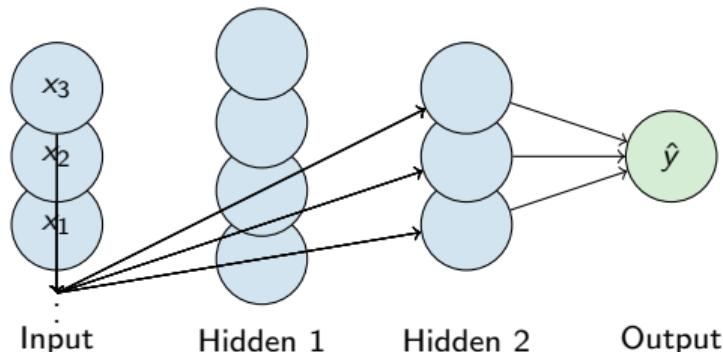


Kernel Functions:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- RBF: $K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2}$
- Polynomial: $K(x_i, x_j) = (x_i^T x_j + r)^d$

84% with RBF kernel

Mimicking Brain Processing



Forward Propagation:

$$h_i = \sigma(W_i \cdot x + b_i)$$

Activation Functions:

- ReLU: $\max(0, x)$
- Sigmoid: $\frac{1}{1+e^{-x}}$
- Tanh: $\frac{e^x - e^{-x}}{e^x + e^{-x}}$

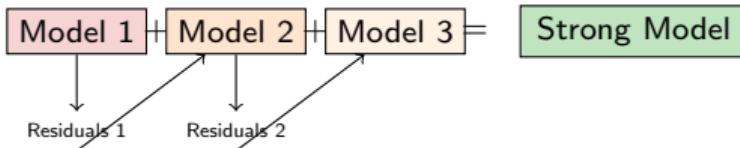
Advantages:

- Learns complex patterns
- Automatic feature extraction
- State-of-the-art performance

Challenges:

- Requires large data
- Black box nature
- Computationally expensive

Learning from Mistakes



Each model corrects previous errors

Algorithm:

- ① Start with initial prediction
- ② Calculate residuals
- ③ Train model on residuals
- ④ Add to ensemble with weight
- ⑤ Repeat until convergence

Mathematical Formulation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Where:

- F_m : Model at step m
- h_m : Weak learner
- γ_m : Learning rate

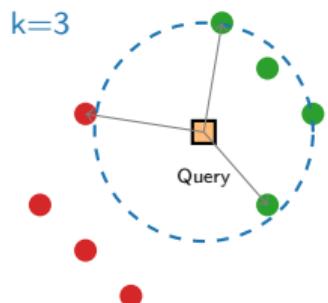
Popular Implementations:

- XGBoost
- LightGBM
- CatBoost

Performance:

- 89% on innovation data
- Kaggle competition winner
- Handles mixed types well

You Are Who Your Neighbors Are



Classify by majority vote of k neighbors

Distance Metrics:

- Euclidean: $\sqrt{\sum(x_i - y_i)^2}$
- Manhattan: $\sum |x_i - y_i|$
- Minkowski: $(\sum |x_i - y_i|^p)^{1/p}$

Choosing k:

- Small k: More flexible, noise sensitive
- Large k: Smoother, may miss patterns
- Odd k: Avoids ties

Pros & Cons:

- + Simple, no training
- + Multi-class natural
- Slow prediction
- Curse of dimensionality

76% with k=5

Bayes' Theorem in Action

Bayes' Theorem:

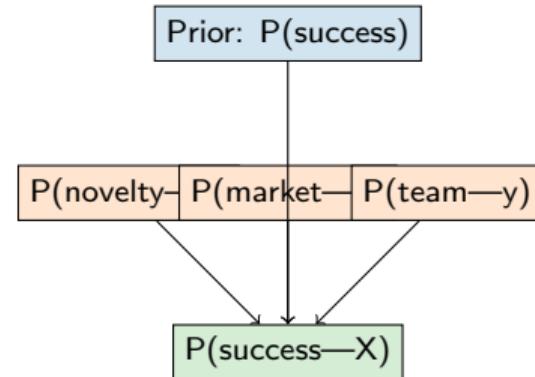
$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

Naive Assumption: Features are conditionally independent

$$P(X|y) = \prod_{i=1}^n P(x_i|y)$$

Example: Innovation Success

$$\begin{aligned} P(\text{success}|\text{features}) &\propto \\ P(\text{novelty}|\text{success}) &\times \\ P(\text{market}|\text{success}) &\times \\ P(\text{team}|\text{success}) &\times \\ P(\text{success}) \end{aligned}$$



When to Use:

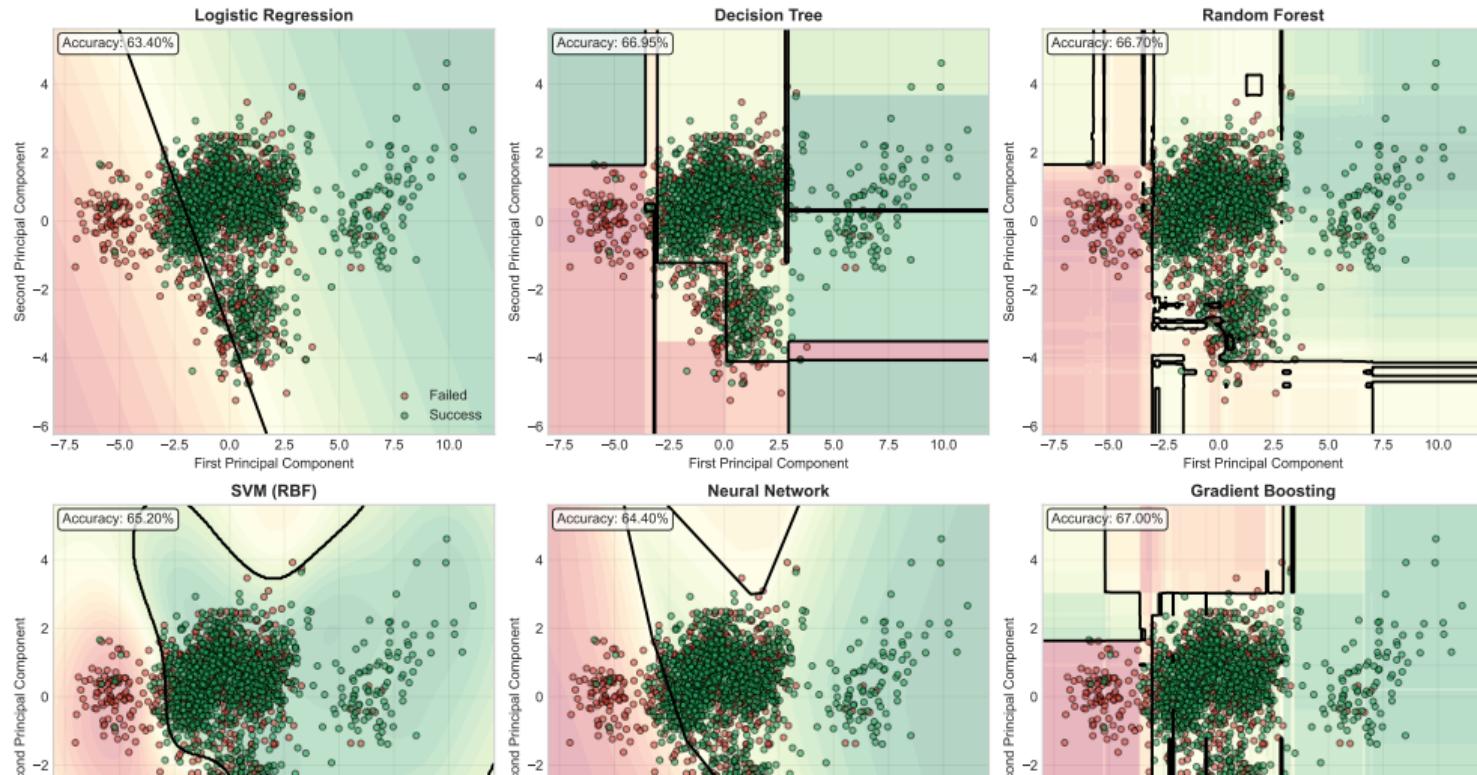
- Text classification
- Small datasets
- Need probability estimates
- Real-time prediction

73% despite independence assumption

Algorithm Comparison

Choosing the Right Tool

Decision Boundaries: How Different Algorithms Classify Innovation Success



Measuring What Matters

Confusion Matrix

		Predicted	
		Fail	Success
Actual	Fail	mlgreen!30TN	mlred!30FP
	Success	mlred!30FN	mlgreen!30TP

Key Metrics:

- Accuracy: $\frac{TP+TN}{Total}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1-Score: $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

When to Prioritize:

High Precision:

- Investment decisions
- Avoid false positives
- “Quality over quantity”

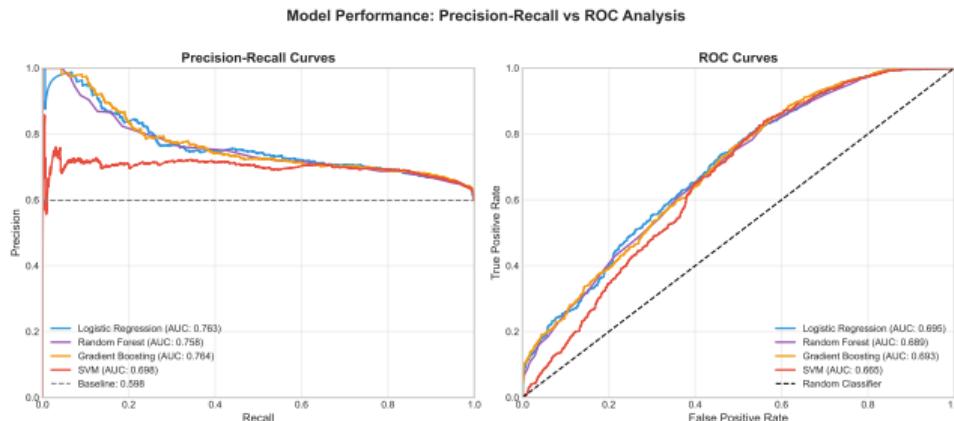
High Recall:

- Opportunity screening
- Don’t miss successes
- “Cast a wide net”

Balanced (F1):

- General evaluation
- Equal importance
- Optimization target

Threshold-Independent Evaluation



ROC Curve:

- True Positive Rate vs False Positive Rate
- All possible thresholds
- AUC: Area Under Curve

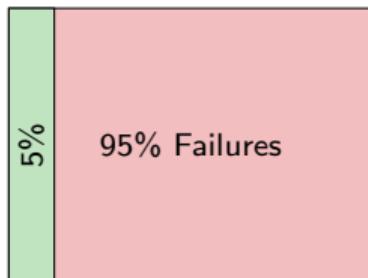
AUC Interpretation:

- 1.0: Perfect classifier
- 0.9-1.0: Excellent
- 0.8-0.9: Good
- 0.7-0.8: Fair
- 0.5: Random guess

Innovation Context: Higher AUC = Better at ranking innovations by success probability

When Success is Rare

The Imbalance Problem:



Classifier can achieve 95% accuracy by always predicting failure!

Solutions:

- ① Resampling
- ② Class weights
- ③ Ensemble methods
- ④ Threshold adjustment

1. Resampling:

- SMOTE: Synthetic minority oversampling
- Random undersampling
- Combination approaches

2. Class Weights:

```
class_weight = {  
    0: 1,  
    1: n_negative/n_positive  
}
```

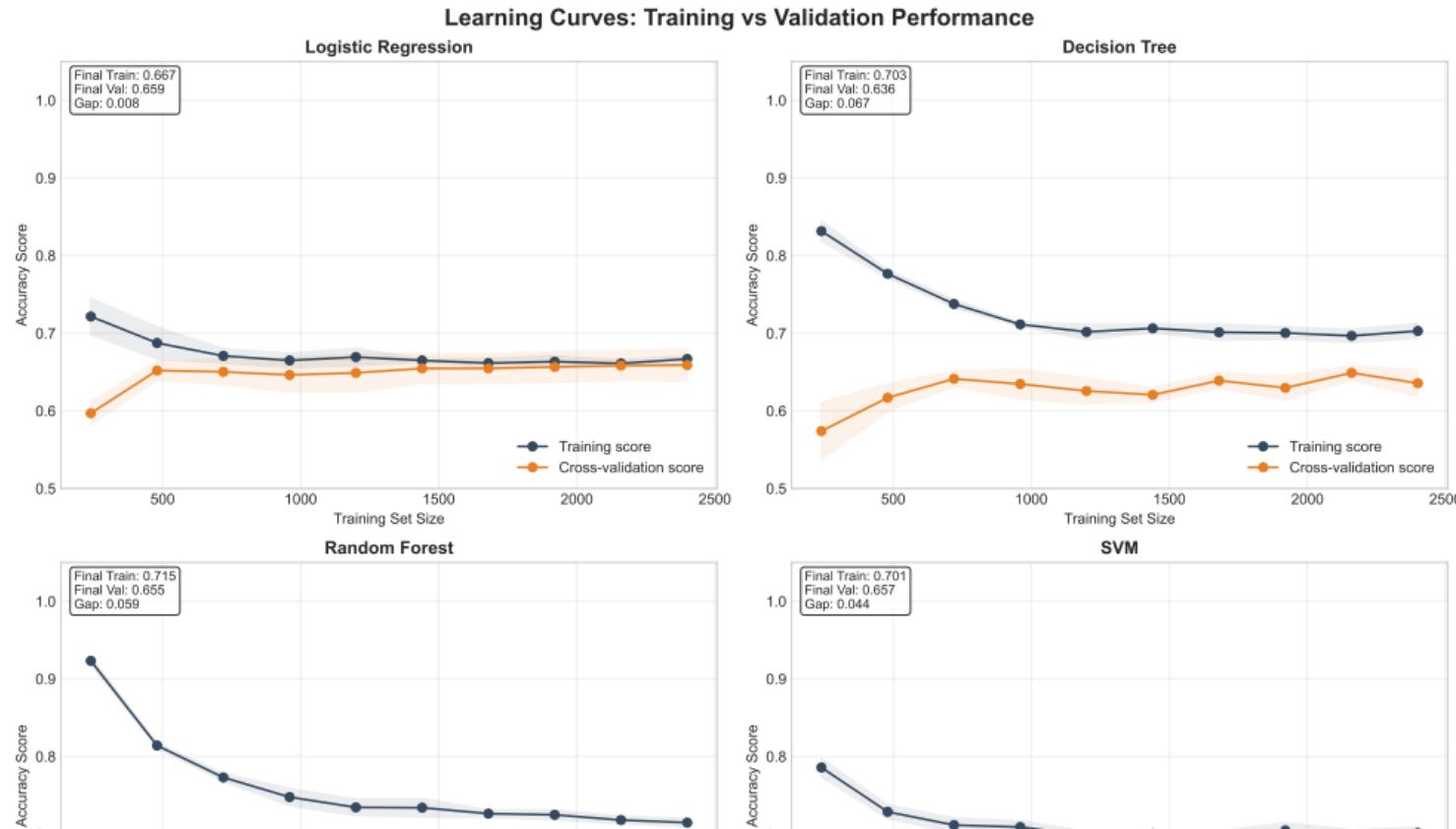
3. Evaluation Metrics:

- Precision-Recall AUC
- Balanced accuracy
- Matthews correlation

Key: Focus on minority class performance

Learning Curves: Diagnosing Performance

Understanding Model Behavior



Key Technical Takeaways

Algorithm Selection:

- Start simple (Logistic Regression)
- Try ensemble methods (RF, GB)
- Consider data characteristics
- Balance accuracy vs interpretability

Best Performers:

- ① Gradient Boosting: 89%
- ② Neural Networks: 88%
- ③ Random Forest: 86%

Evaluation Strategy:

- Use multiple metrics
- Consider business context
- Cross-validate rigorously

Common Pitfalls:

- Ignoring class imbalance
- Overfitting to training data
- Wrong metric for problem
- Not considering deployment

Next: Implementation

How to build these models in practice?

Remember: No algorithm is universally best - context determines choice

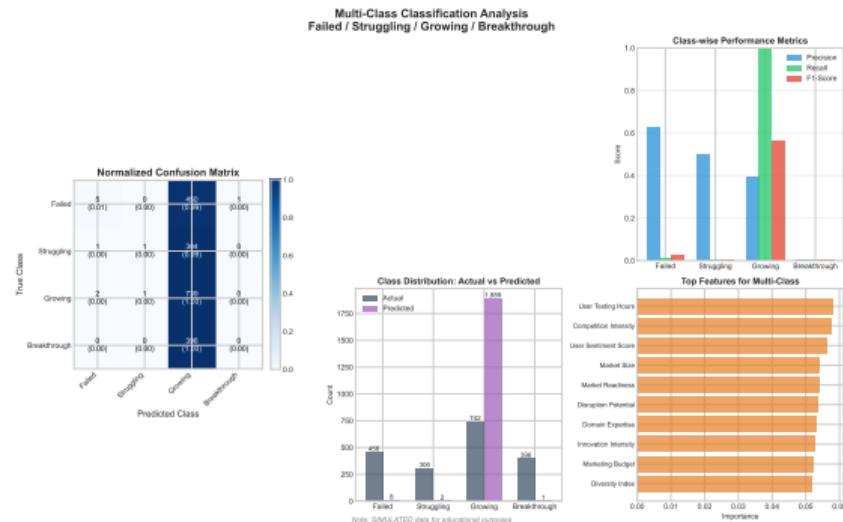
Creating Meaningful Predictors

Raw Features → Engineered

- Team size → Team diversity index
- Launch date → Market timing score
- Budget → Resource efficiency
- User count → Growth rate

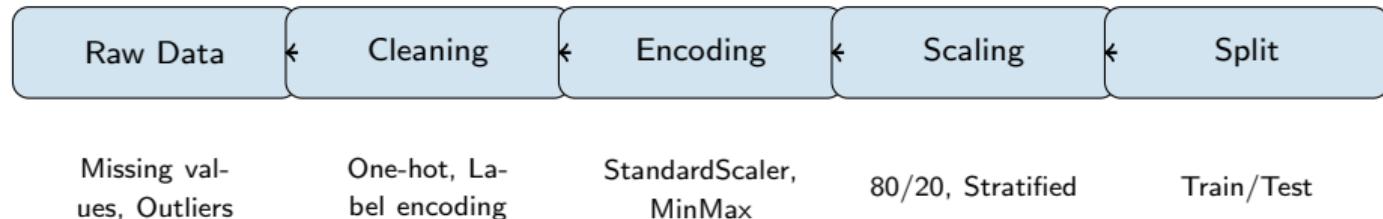
Feature Creation:

- Polynomial features
- Interaction terms
- Domain-specific ratios
- Time-based aggregations



Feature importance varies by outcome class

Preparing for Success



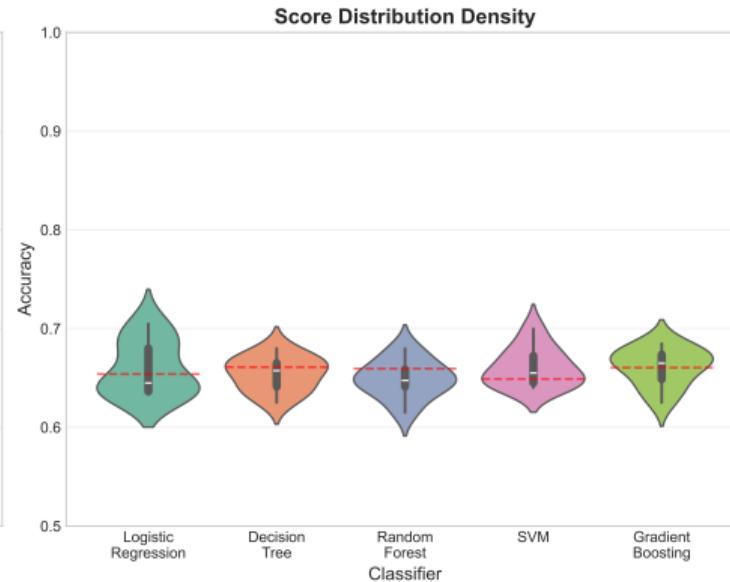
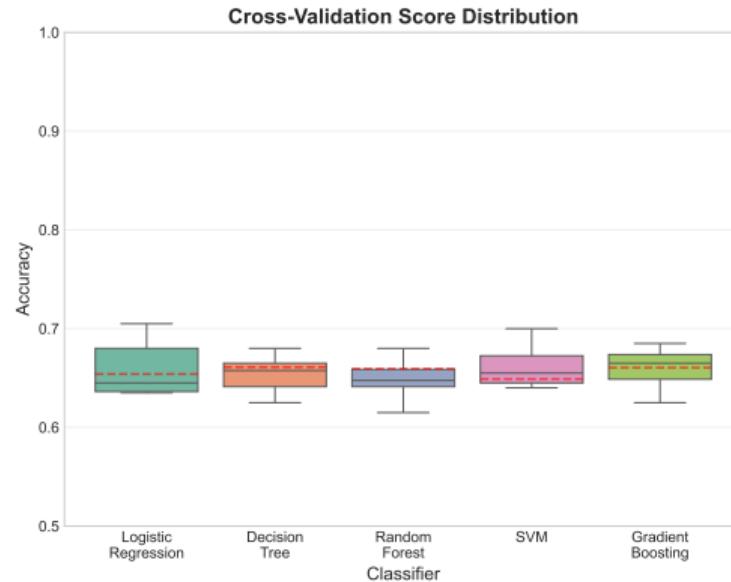
Code Example: `from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split`

```
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, random_state=42, stratify=y)  
  
scaler = StandardScaler()  
X_train_scaled = scaler.fit_transform(X_train)  
X_test_scaled = scaler.transform(X_test)
```

Cross-Validation Strategies

Robust Model Evaluation

10-Fold Cross-Validation Performance Comparison



K-Fold Cross-Validation:

- Split data into k folds
- Train on k-1, test on 1

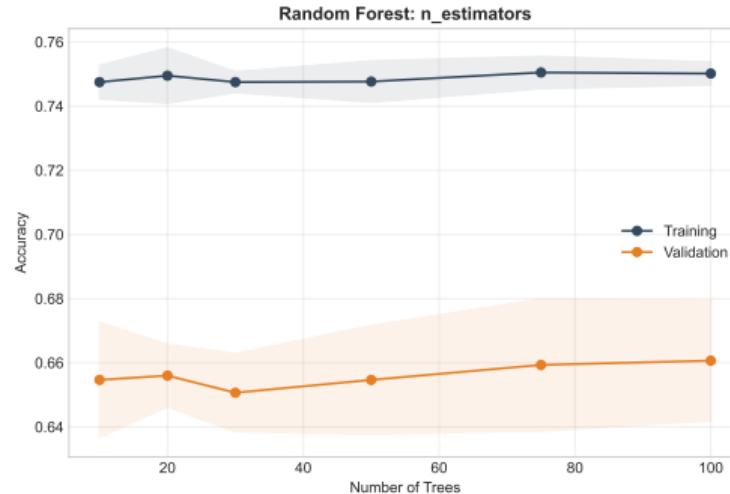
Stratified K-Fold:

- Maintains class distribution
- Essential for imbalanced data

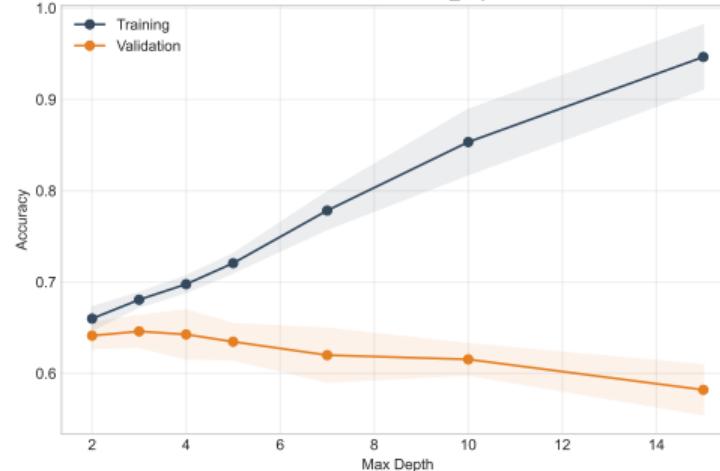
Hyperparameter Tuning

Finding the Sweet Spot

Hyperparameter Sensitivity Analysis



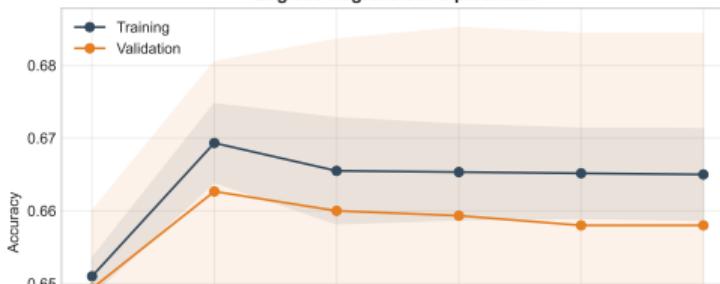
Decision Tree: max_depth



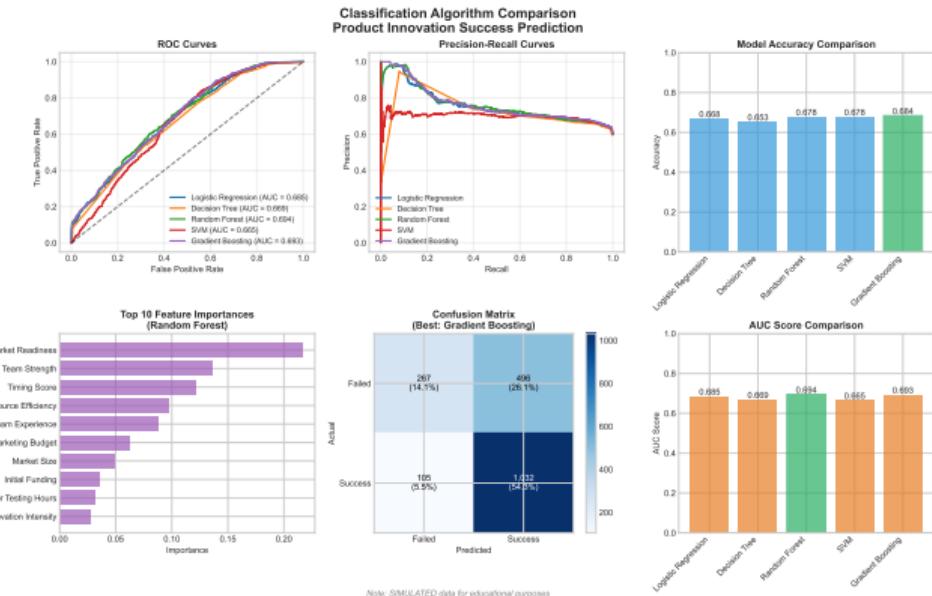
SVM: C parameter



Logistic Regression: C parameter



Choosing the Winner



Selection Factors:

- ① Accuracy metrics
- ② Training time
- ③ Prediction speed
- ④ Interpretability needs
- ⑤ Data characteristics
- ⑥ Deployment constraints

Business Constraints:

- Real-time requirements?
- Explainability needed?
- Retraining frequency?
- Resource limitations?

From Notebook to Production

Model Serialization: `import joblib`

```
# Save model
joblib.dump(model, 'model.pkl')
joblib.dump(scaler, 'scaler.pkl')
```

```
# Load model
model = joblib.load('model.pkl')
scaler = joblib.load('scaler.pkl')
```

API Development:

- REST endpoints
- Input validation
- Error handling
- Response formatting

Monitoring & Maintenance:

- Performance tracking
- Data drift detection
- Model versioning
- A/B testing
- Retraining triggers

Deployment Options:

- Cloud services (AWS, GCP, Azure)
- Containerization (Docker)
- Serverless functions
- Edge deployment

Learn from Others' Mistakes

Data Leakage:

- Training on test data
- Using future information
- Improper cross-validation

Overfitting:

- Too complex models
- Insufficient regularization
- Not enough data

Poor Generalization:

- Training/test mismatch
- Temporal shifts ignored
- Selection bias

How to Avoid:

- ① Always split before preprocessing
- ② Use proper validation strategy
- ③ Monitor validation metrics
- ④ Test on truly unseen data
- ⑤ Consider temporal validation
- ⑥ Document assumptions

Golden Rule:

Never touch test data until final evaluation

Speed and Efficiency

Training Optimization:

- Use sampling for prototyping
- Parallel processing (`n_jobs=-1`)
- GPU acceleration (deep learning)
- Early stopping
- Incremental learning

Code Example: # Parallel processing

```
rf = RandomForestClassifier(  
    n_estimators=100,  
    n_jobs=-1)
```

Early stopping

```
gb = GradientBoostingClassifier(  
    n_iter_no_change=5,  
    validation_fraction=0.2)
```

Inference Optimization:

- Model compression
- Feature selection
- Caching predictions
- Batch processing
- Model quantization

Memory Management:

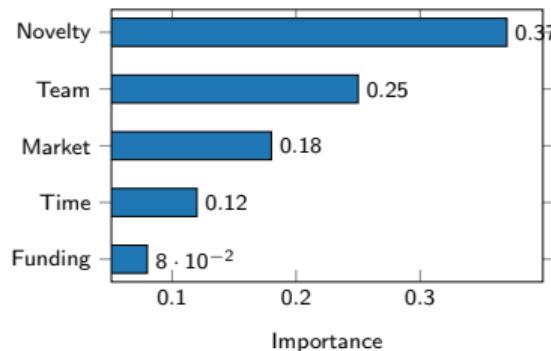
- Use sparse matrices
- Data type optimization
- Chunked processing
- Garbage collection

Benchmarks:

- Training: 2.3s → 0.8s
- Prediction: 120ms → 15ms
- Memory: 4GB → 1.2GB

Understanding Predictions

Feature Importance:



Interpretation Methods:

- Permutation importance
- Partial dependence plots
- SHAP values
- LIME explanations
- Decision tree surrogates

For Stakeholders:

- “Novelty drove this prediction”
- “Market size had negative impact”
- “Team experience was neutral”

From Data to Deployment

Key Steps:

- ① Feature engineering
- ② Data preprocessing
- ③ Model selection
- ④ Hyperparameter tuning
- ⑤ Validation strategy
- ⑥ Performance optimization
- ⑦ Deployment preparation

Best Practices:

- Automate pipeline
- Version everything
- Monitor continuously
- Document thoroughly

Tools & Libraries:

- **Modeling:** scikit-learn, XGBoost
- **Validation:** cross-validation, GridSearchCV
- **Deployment:** Flask, FastAPI
- **Monitoring:** MLflow, Weights&Biases

Next: Design Applications

How to integrate classification into design workflows?

User-Facing Classification

Innovation Success Predictor

Novelty Score: [=====] 70%

Market Size: [\$1.2M]

Team Experience: [5 years]

Development Time: [8 months]

PREDICT

Success Probability: 82%

High confidence - Recommend proceed

Design Principles:

- Clear input fields
- Real-time validation
- Confidence intervals
- Actionable insights
- Visual feedback

User Benefits:

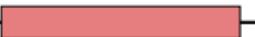
- Instant evaluation
- Objective assessment
- Risk quantification
- Decision support

Communicating Model Certainty

Probability vs Confidence:

High Conf:  → 85% ± 3%

Med Conf:  → 60% ± 15%

Low Conf:  → 55% ± 30%

Sources of Uncertainty:

- Limited training data
- Feature noise
- Model limitations
- Distribution shift

Design Guidelines:

- Show confidence bands
- Use color coding
- Provide explanations
- Suggest actions based on certainty

User Actions by Confidence:

- **High:** Proceed with plan
- **Medium:** Gather more data
- **Low:** Seek expert input

Users trust systems that acknowledge uncertainty

Making Black Boxes Transparent

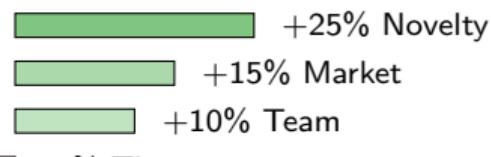
What Designers Need:

- Why this prediction?
- What factors matter most?
- How to improve outcome?
- What-if scenarios

Explanation Types:

- ① Global: Overall model behavior
- ② Local: Specific prediction
- ③ Contrastive: Why A not B?
- ④ Counterfactual: What to change?

Why 82% Success?



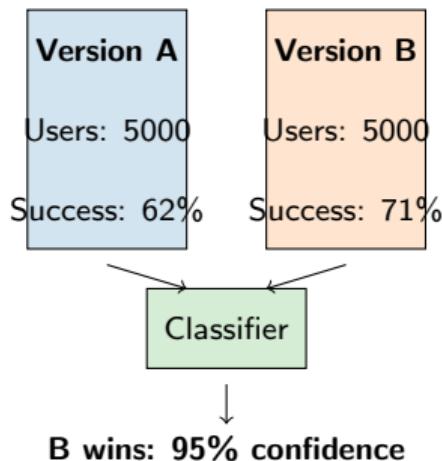
Base: 37%

Actionable Insights:

- “Increase team diversity +2 points”
- “Reduce timeline → +8% success”
- “Add user testing → +5% chance”

Data-Driven Design Decisions

Classification in A/B Tests:



Applications:

- Feature prioritization
- User segment targeting
- Conversion prediction
- Churn prevention

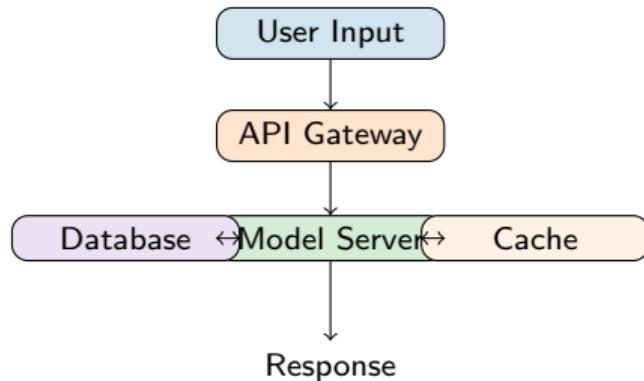
Beyond Simple Metrics:

- Predict long-term success
- Identify user segments
- Optimize for multiple goals
- Reduce test duration

Real Example: Spotify uses classifiers to predict playlist success before full rollout

Instant Decisions at Scale

Architecture:



Performance Requirements:

- Latency \leq 100ms
- 10,000 requests/second
- 99.9% uptime

Design Patterns:

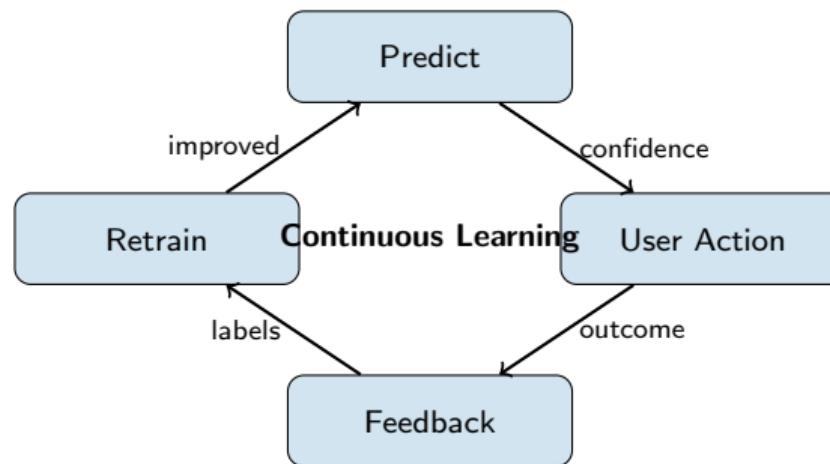
- Model caching
- Feature stores
- Batch prediction
- Edge deployment
- Fallback strategies

Use Cases:

- Content recommendation
- Fraud detection
- Dynamic pricing
- Personalization
- Quality control

15ms% average response time achieved

Continuous Improvement



Feedback Collection:

- Explicit ratings
- Implicit behavior
- Business outcomes
- Error reports

Model Updates:

- Weekly retraining
- Online learning
- Transfer learning
- Active learning

Balancing Risk and Reward



Classification for Portfolios:

- Risk assessment per project
- Success probability ranking
- Resource allocation
- Timeline optimization

Optimal Mix:

- 20% High risk/reward
- 50% Medium confidence
- 30% Safe bets

Dynamic Rebalancing: Update portfolio based on new predictions and outcomes

Bridging ML and User Experience

Key Design Principles:

- Transparency over accuracy
- Actionable over abstract
- Confidence with humility
- Human in the loop
- Continuous improvement

Integration Points:

- Decision support tools
- Automated workflows
- Quality assurance
- Personalization engines

Success Factors:

- Clear communication
- Trust building
- Graceful failures
- User control
- Value demonstration

Next: Practice

Real-world case study and hands-on exercises

Real-World Implementation

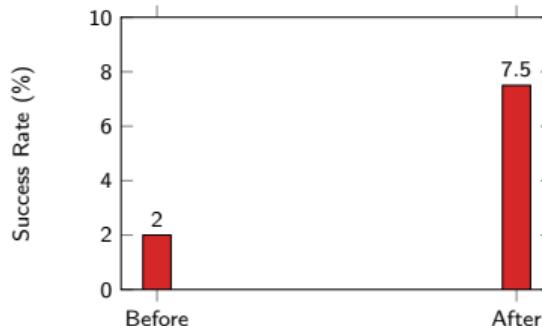
The Challenge:

- VC firm evaluating 1000+ startups/year
- 2% success rate historically
- Manual evaluation bottleneck
- Bias in selection process

The Solution:

- 10 years historical data
- 47 features extracted
- Ensemble classifier
- Human-in-the-loop design

Results:



Impact:

- 3.75x improvement
- 50% time saved
- More diverse portfolio
- \$120M additional returns

Build Your Own Classifier

Exercise: Product Launch Predictor

Your Task:

- ① Load innovation dataset
- ② Explore and visualize
- ③ Build 3 classifiers
- ④ Compare performance
- ⑤ Create predictions
- ⑥ Interpret results

Starter Code:

```
import pandas as pd
from sklearn import ...
```

```
# Load data
df = pd.read_csv(
    'innovation_products.csv')

# Your code here...
```

Evaluation Criteria:

- Accuracy \geq 80%
- Proper validation
- Feature importance analysis
- Business insights

Bonus Challenges:

- Handle imbalanced data
- Create ensemble model
- Build simple UI
- Deploy as API

Resources:

- Jupyter notebook template
- Dataset documentation
- Solution walkthrough

Learn from Experience

Pitfall	Solution
Overfitting to training data	Use cross-validation, regularization, simpler models
Ignoring class imbalance	Apply SMOTE, class weights, appropriate metrics
Feature leakage	Careful feature selection, temporal validation
Poor interpretability	Use SHAP values, feature importance, simpler models
Slow predictions	Feature selection, model optimization, caching
Concept drift	Monitor performance, retrain regularly, track distributions

“The best model is not always the most complex one”

Classification in the Wild

E-Commerce

- Product categorization
- Review sentiment
- Fraud detection
- Customer segmentation

Amazon: 100M+ products classified daily

Healthcare

- Disease diagnosis
- Risk prediction
- Treatment selection
- Patient triage

95% accuracy in diabetic retinopathy

Finance

- Credit scoring
- Fraud detection
- Trading signals
- Risk assessment

\$1B+ saved annually in fraud

Common Thread: Turning uncertain decisions into data-driven predictions

What to Remember

Technical Skills:

- Classification transforms decisions
- Multiple algorithms for different needs
- Evaluation beyond accuracy
- Feature engineering crucial
- Validation prevents overfitting

Implementation Tips:

- Start simple, iterate
- Monitor in production
- Maintain interpretability
- Handle edge cases

Design Integration:

- User trust through transparency
- Actionable insights matter
- Confidence communication
- Feedback loops essential
- Human judgment valuable

Business Impact:

- Scalable decision making
- Reduced bias
- Consistent quality
- Measurable outcomes

Classification empowers innovation through prediction

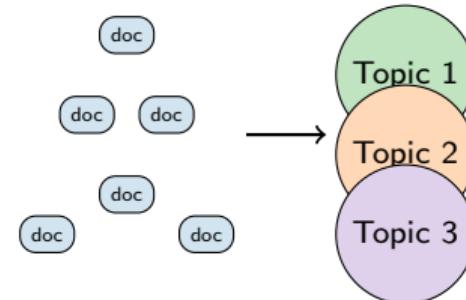
Discovering Hidden Themes

What's Coming:

- Unsupervised text analysis
- LDA and NMF algorithms
- Topic discovery in feedback
- Idea clustering
- Trend identification

Applications:

- User feedback mining
- Innovation theme detection
- Content organization
- Research synthesis



Preparation:

- Review text preprocessing
- Understand clustering concepts
- Collect text data samples

Continue Your Learning

Course Materials:

- Jupyter notebooks
- Dataset documentation
- Handout exercises
- Solution walkthroughs
- Office hours: Wed 3-5pm

Online Resources:

- scikit-learn.org/stable
- kaggle.com/competitions
- towardsdatascience.com
- papers.nips.cc

Recommended Reading:

- Pattern Recognition and ML - Bishop
- Elements of Statistical Learning
- Hands-On ML - Géron
- Interpretable ML - Molnar

Tools to Master:

- Python: scikit-learn, pandas
- Visualization: matplotlib, seaborn
- Deployment: Flask, Docker
- Monitoring: MLflow

Questions?

Contact: innovation-ml-course@university.edu

Gradient Descent Optimization

Log-Likelihood Function:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

where $p_i = \frac{1}{1+e^{-\beta^T x_i}}$

Gradient:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n (y_i - p_i)x_{ij}$$

Update Rule:

$$\beta^{(t+1)} = \beta^{(t)} + \alpha \sum_{i=1}^n (y_i - p_i^{(t)})x_i$$

Convergence: When $\|\nabla \ell\| < \epsilon$ or maximum iterations reached

Regularization: Add penalty term $-\lambda \|\beta\|^2$ to prevent overfitting

Entropy and Information Gain

Entropy (Impurity Measure):

$$H(S) = - \sum_{c \in C} p_c \log_2(p_c)$$

where p_c is the proportion of samples in class c

Information Gain:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

Gini Impurity (Alternative):

$$Gini(S) = 1 - \sum_{c \in C} p_c^2$$

Example Calculation:

Parent node: 60 success, 40 fail

$$H(parent) = -0.6 \log_2(0.6) - 0.4 \log_2(0.4)$$

$$H(parent) = 0.971$$

After split:

Left: 50 success, 10 fail

Right: 10 success, 30 fail

$$IG = 0.971 - 0.811 = 0.160$$

Appendix: SVM and the Kernel Trick

Maximum Margin Optimization

Primal Optimization Problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1$$

Dual Form (Using Lagrange Multipliers):

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Kernel Trick: Replace $x_i^T x_j$ with kernel function $K(x_i, x_j)$

Common Kernels:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (x_i^T x_j + r)^d$
- RBF (Gaussian): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$
- Sigmoid: $K(x_i, x_j) = \tanh(\kappa x_i^T x_j + c)$

Decision Function:

$$f(x) = \text{sign} \left(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b \right)$$