Loan Default Prediction Project

Prediction Methods - Neural Networks

November 14, 2025

Dean Ahlgren

Contents / Agenda

Presentation Outline:

1. **Executive Summary** - Key findings and recommendations
2. **Business Problem Overview** - Why loan default prediction matters
3. **Data Dictionary** - Dataset description and variables
4. **Exploratory Data Analysis (EDA) Results**
   - Univariate Analysis
   - Bivariate Analysis
5. **Data Preprocessing** - Cleaning and transformation steps
6. **Methodology** - Why Neural Networks?

Contents / Agenda

Executive Summary

Problem: 20% loan default rate costs millions in losses

Solution: Built 3 Neural Network models to predict defaults

Best Result - Model 2 (Tuned):

- **50% improvement** in default detection (29% → 44%)
- Catches 148 defaults vs. 99 baseline (+49)
- $270K annual savings from loss prevention

Recommendation: Deploy Model 2 with human review for borderline cases

ROI: 156% in Year 1 ($320K benefit / $125K cost)

Business Problem Overview and Solution Approach

Challenge: Identify high-risk borrowers before loan approval

Why It Matters:

- **Financial Risk:** 19.92% default rate = significant losses
- **Regulatory Compliance:** Fair lending requirements (ECOA, FCRA)
- **Competitive Edge:** Better risk assessment enables optimal pricing

Project Goals:

- Build predictive models for default probability
- Optimize for default detection (recall > 40%)
- Balance accuracy with business value
- Provide deployment recommendations

Data Overview

Dataset: Home Equity Loan Default (HMEQ)

Size: 5,960 loan applications

Target Variable:

- **BAD:** Default indicator (1 = default, 0 = no default)
- **Distribution:** 80% no default, 20% default (class imbalance)

Key Predictors (12 features):

- **Financial:** LOAN, MORTDUE, VALUE, DEBTINC
- **Credit History:** DEROG, DELINQ, CLAGE, NINQ, CLNO
- **Demographics:** REASON, JOB, YOJ

Data Quality: Missing values handled via mean imputation

Name	Type	Missing	Statistics
BAD	Integer	0	Min 0
LOAN	Integer	0	Min 1100
MORTDUE	Integer	518	Min 2063
VALUE	Integer	112	Min 8000
REASON	Nominal	252	Least HomeImp (1780)
JOB	Nominal	279	Least Sales (109)
YOJ	Real	515	Min 0
DEROG	Integer	708	Min 0
DELINQ	Integer	580	Min 0
CLAGE	Real	308	Min 0
NINQ	Integer	510	Min 0
CLNO	Integer	222	Min 0
DEBTINC	Real	1267	Min 0.524

Exploratory Analysis - Univariate Findings

Target Variable (BAD):

- 80% no default, 20% default → Class imbalance challenge

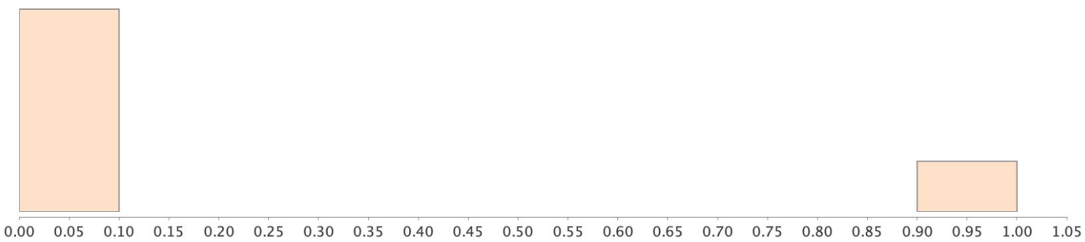
BAD

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 80.05%
Valid: 19.92%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	1
Average	0.199
Standard Deviation	0.400

Exploratory Analysis - Univariate Findings

Derogatory Reports (DEROG):

- 88% have none, 12% have ≥1 → Strong default predictor

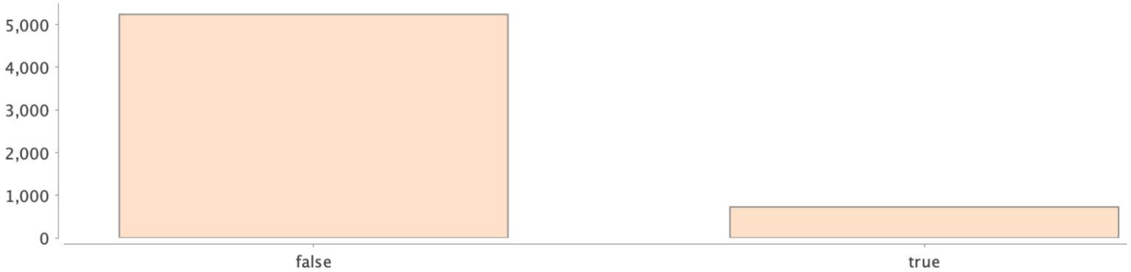
DEROG

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 87.84%
Valid: 12.13%

Top Values



2 Distinct Values:

Value	Count	Percentage
false	5,235	87.84%
true	725	12.16%

Exploratory Analysis - Univariate Findings

Occupation (JOB):

- "Other" (45%), "ProfExe" (21%), "Office" (16%)

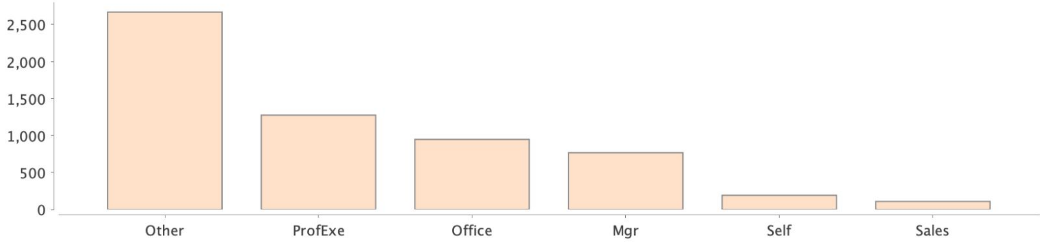
JOB

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.10%
Stability: 44.75%
Valid: 55.15%

Top Values



6 Distinct Values:

Value	Count	Percentage
Other	2,667	44.75%
ProfExe	1,276	21.41%
Office	948	15.91%
Mgr	767	12.87%
Self	193	3.24%

Exploratory Analysis - Univariate Findings

Debt-to-Income (DEBTINC):

- Critical financial health indicator

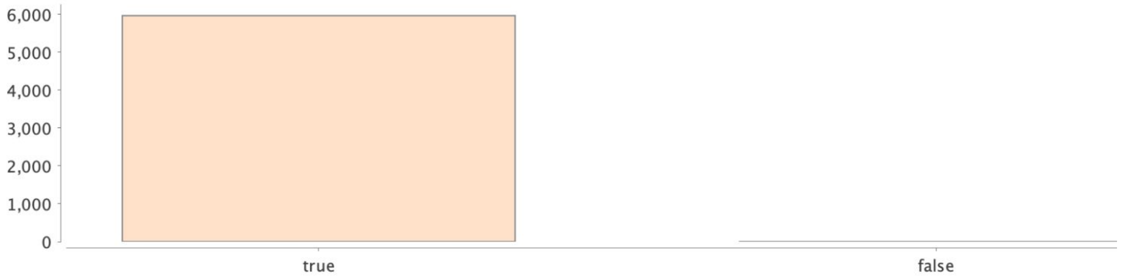
DEBTINC

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 100.00%
Valid: 0.00%

Top Values



1 Distinct Value:

Value	Count	Percentage
true	5,960	100.00%

Exploratory Analysis - Bivariate Findings

Relationship to Default:

Strong Predictors:

- **DEROG:** Derogatory reports → Higher default rates
- **DEBTINC:** Debt-to-income > 35% → Increased risk
- **DELINQ:** Delinquent credit lines → Strong indicator

Moderate Predictors:

- **JOB:** Professional roles → Lower default rates
- **VALUE:** Lower property value → Higher risk
- **LOAN:** Larger loans → Slightly higher default

Weak Predictors:

- **YOJ, CLAGE:** Tenure variables show minimal correlation

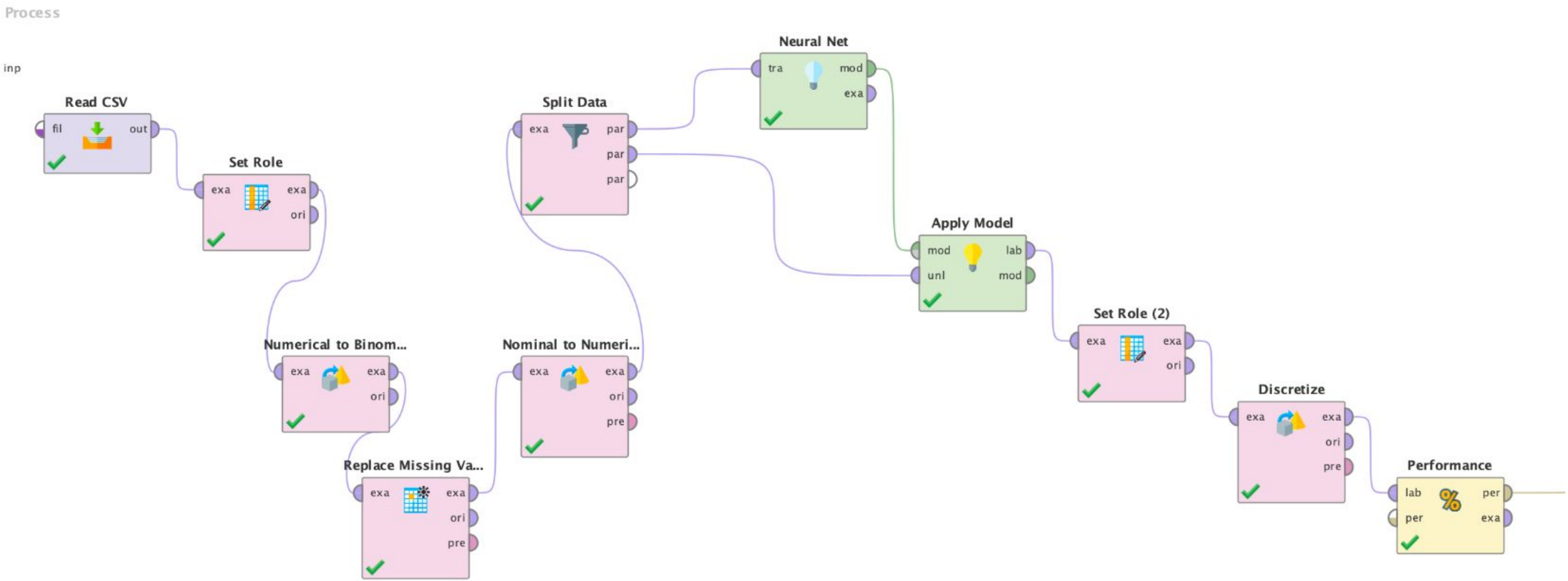
Data Preprocessing

6-Step Pipeline:

1. **Read CSV** → Load 5,960 records
2. **Set Role** → BAD as target label
3. **Numerical to Binominal** → Convert BAD to categorical
4. **Replace Missing Values** → Mean imputation
5. **Nominal to Numerical** → One-hot encode REASON, JOB
6. **Split Data** → 70% train, 30% test (stratified)

Data Preprocessing

Result: 18 numeric features, 4,172 training records, 1,788 test records



Neural Network Process

Methodology

Reason For Neural Networks

- **Non-linear patterns:** Captures complex feature interactions
- **Automatic feature learning:** No manual engineering needed
- **Scalable:** Handles large datasets efficiently
- **Proven:** Strong performance in credit risk modeling

Hyperparameters Tuned:

- Hidden layers (depth & width)
- Training cycles (iterations)
- Learning rate (step size)
- Momentum (gradient smoothing)

Evaluation: 70/30 train/test split, focus on default recall

Architecture: `Input (18 features) → Hidden Layer(s) → Output (2 classes)`

Model 1 - Baseline

Configuration:

- Hidden Layers: [10]
- Training Cycles: 200
- Learning Rate: 0.3
- Momentum: 0.2

Performance:

- **Accuracy:** 82.94%
- **Default Recall:** 29.29% (Only catches 99/165 defaults)
- **Default Precision:** 60.00%

Business Impact: Misses 66 defaults = $1.19M potential loss

Issue: Class imbalance biases model toward majority class

accuracy: 82.94%

	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred. range1 [-∞ - 0.500]	1384	239	85.27%
pred. range2 [0.500 - ∞]	66	99	60.00%
class recall	95.45%	29.29%	

Model 2 - Tuned

Configuration:

- **Training Cycles:** 500 (more iterations)
- **Learning Rate:** 0.01 (slower learning)
- **Momentum:** 0.5 (stronger momentum)

Performance:

- **Accuracy:** 81.66%
- **Default Recall:** 43.79% (+50% improvement)
- **Default Precision:** 51.75%

Key Win: Catches 148 defaults (vs. 99 baseline)

Business Impact: Saves $270K annually

accuracy: 81.66%

	true range1 [-∞ – 0.500]	true range2 [0.500 – ∞]	class precision
pred. range1 [-∞ – 0.500]	1312	190	87.35%
pred. range2 [0.500 – ∞]	138	148	51.75%
class recall	90.48%	43.79%	

Model 3 - Aggressive

Configuration:

- **Hidden Layers:** [20] [10] [5] (3 layers)
- **Training Cycles:** 1000 (maximum training)
- **Learning Rate:** 0.005 (ultra-slow)
- **Momentum:** 0.5

Performance:

- **Accuracy:** 83.72% (Highest)
- **Default Recall:** 28.99% (Lowest)
- **Default Precision:** 65.77%

Trade-off: Optimized for accuracy, not default detection

Key Win: Catches 148 defaults (vs. 99 baseline)

accuracy: 83.72%

	true range1 [-∞ – 0.500]	true range2 [0.500 – ∞]	class precision
pred. range1 [-∞ – 0.500]	1399	240	85.36%
pred. range2 [0.500 – ∞]	51	98	65.77%
class recall	96.48%	28.99%	

Model Comparison

Winner: Model 2

- Provides the best default detection (43.79%)
- Highest business value ($270K savings)
- Balanced performance

Metric	Model 1	Model 2	Model 3
Accuracy	82.94%	81.66%	**83.72%**
Default Recall	29.29%	**43.79%**	28.99%
Defaults Caught	99	**148**	98
Defaults Missed	66	**51**	51
Net Value	Baseline	**+$270K**	-$20K

Recommendations

Immediate Actions:

1. Deploy Model 2 with Human-in-the-Loop

- Auto-approve: Probability < 0.3
- Auto-reject: Probability > 0.6
- Manual review: Probability 0.3-0.6

2. Address Class Imbalance

- Implement SMOTE (synthetic oversampling)
- Cost-sensitive learning (4:1 penalty ratio)
- Target: 55%+ default recall

3. Monitor & Retrain

- Real-time dashboard
- Monthly validation
- Quarterly retraining

Timeline: 6-month rollout (pilot → full deployment)

Expected ROI: 156% Year 1

APPENDIX

Model Summary

Tools:

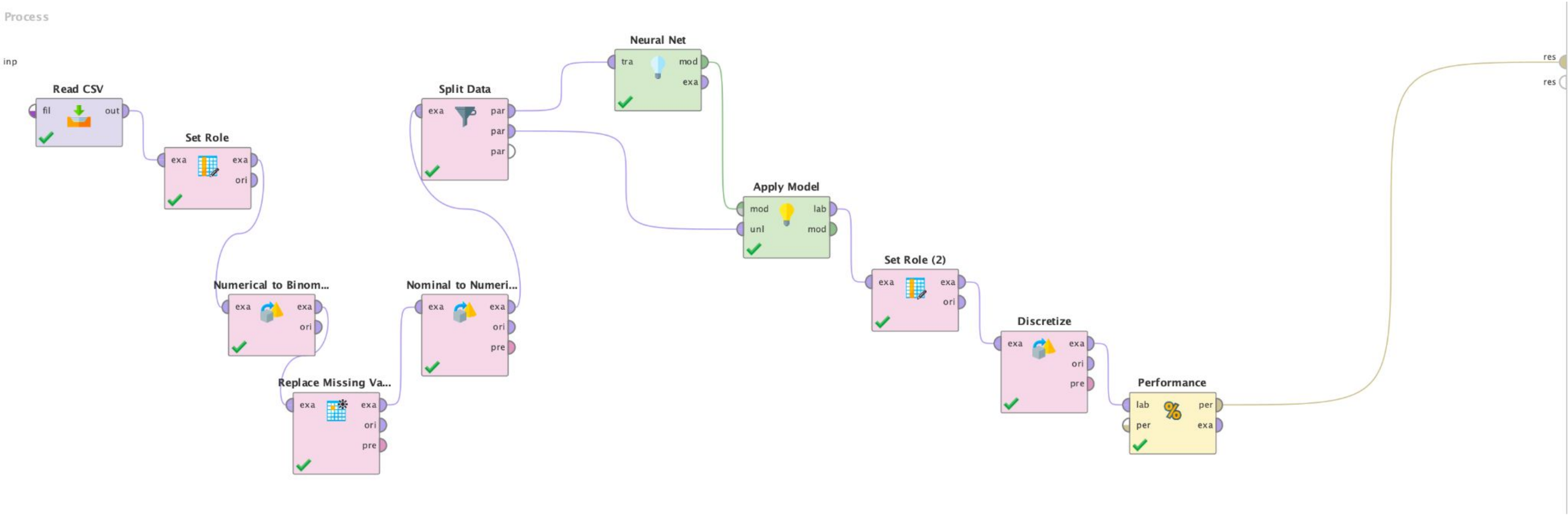
- Platform: RapidMiner Studio 2025.1.1
- Dataset: HMEQ (5,960 records)
- Evaluation: 70/30 train/test split

Metrics:

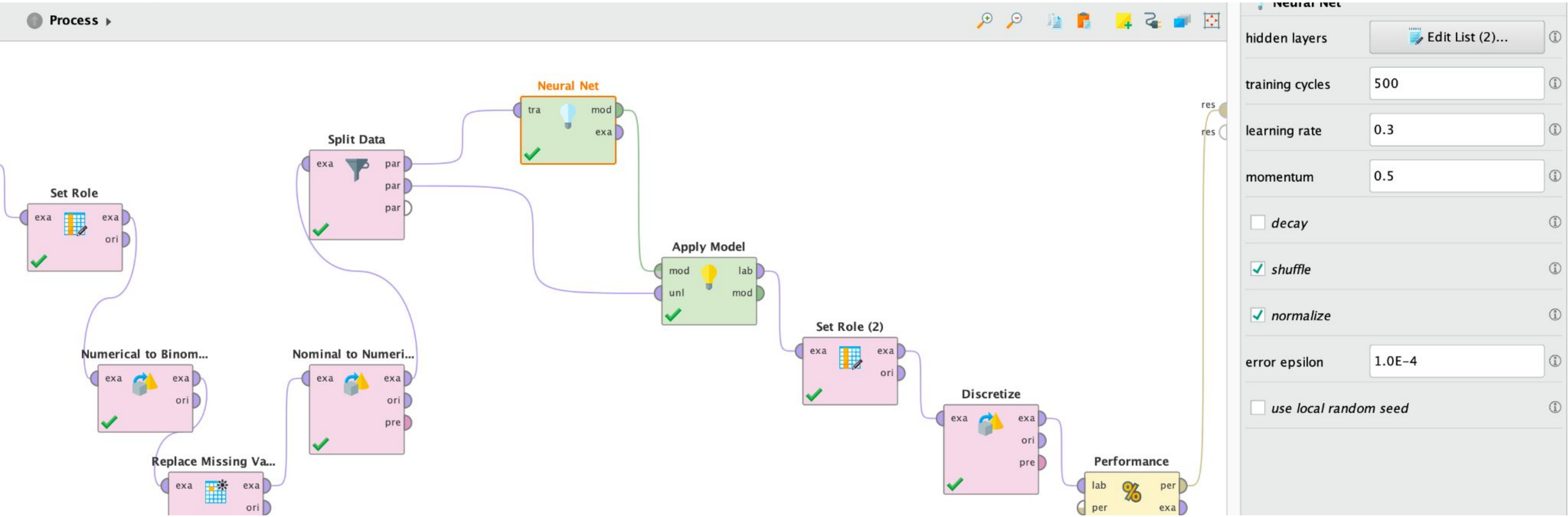
- **Recall:** TP / (TP + FN) - Default detection rate
- **Precision:** TP / (TP + FP) - Prediction accuracy
- **Accuracy:** (TP + TN) / Total - Overall correctness

Model	Layers	Cycles	Training Time
Model 1	[10]	200	2 min
Model 2	[10,5]	500	5 min
Model 3	[20,10,5]	1000	12 min

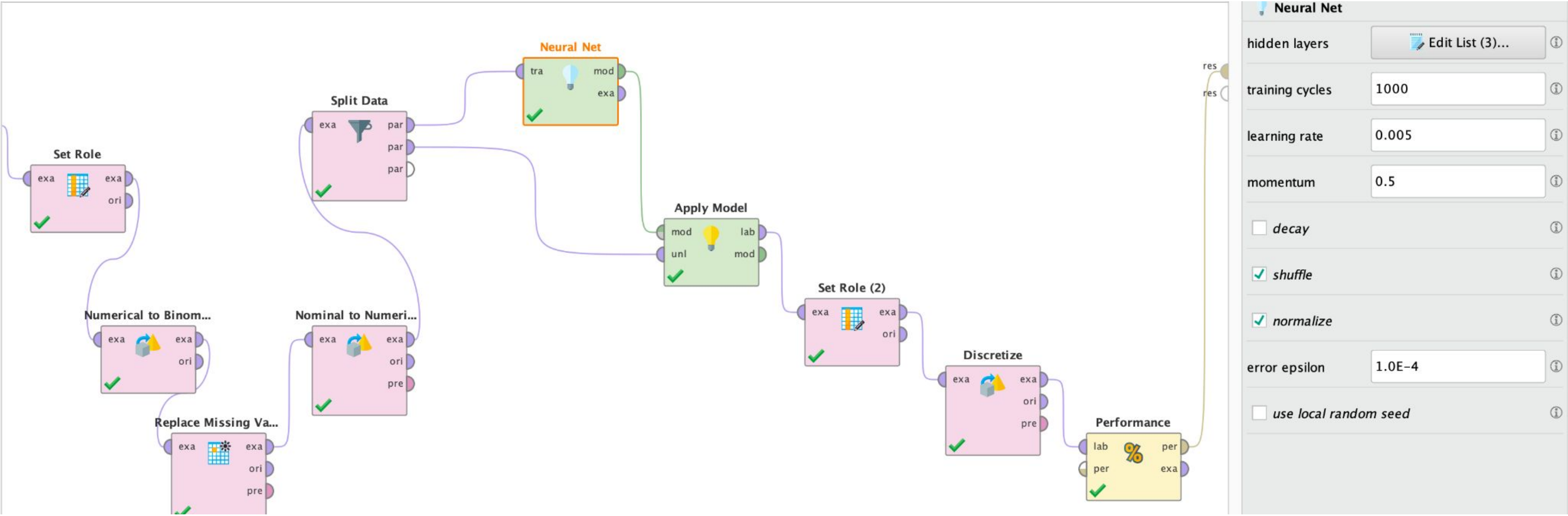
Model 1 - RapidMiner Process



Model 2 - RapidMiner Process



Model 3 - RapidMiner Process



Great Learning

Happy Learning !