Machine Learning for Business Analytics

Prem Kumar Chimakurthi

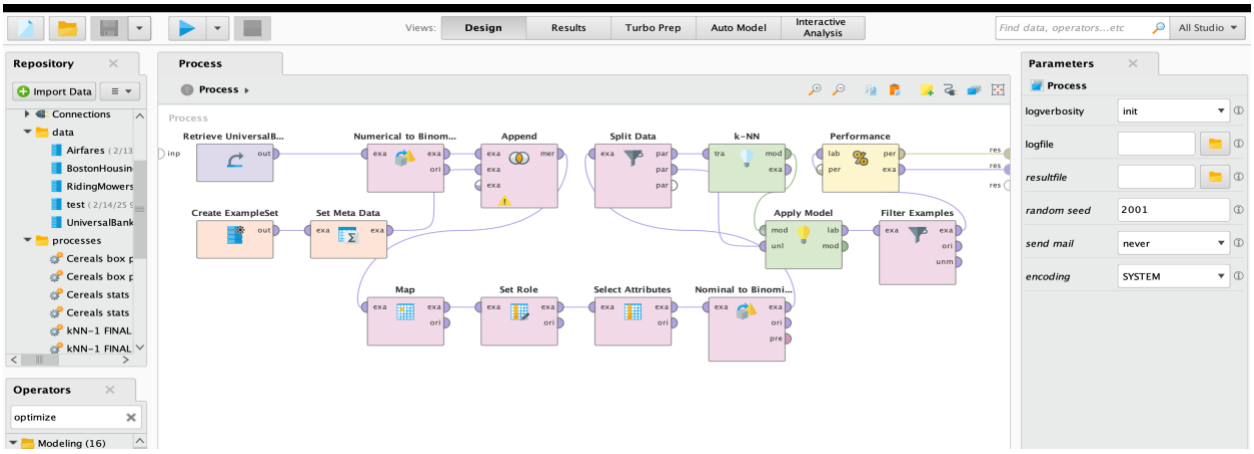
Master of Science in Business Analytics

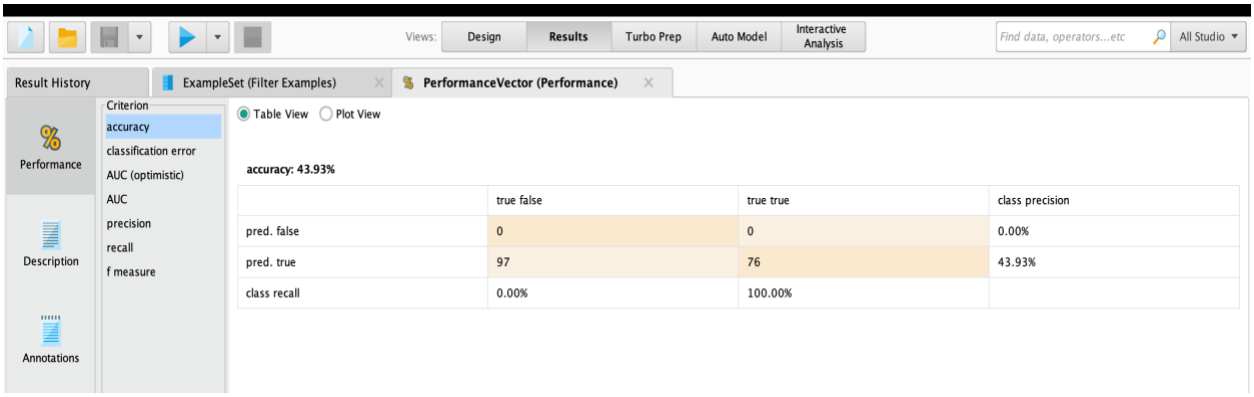
Personal Loan Acceptance Project

Consider the following customers: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education = 2, Mortgage = 0, Securities Account = false, CD Account = false, Online = true, and Credit Card = true.

Perform a k-NN classification with the selected predictors using k=1. Use the default threshold value of 0.5. How would this customer be classified?

Solution:





	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	97	76	43.93%
class recall	0.00%	100.00%	

Nearest Neighbor Analysis: The customer was compared to the closest data point in the training set using k=1.

Threshold Used: The classification was made using the default threshold of 0.5.

Classification Result: The model classified the customer as "true", predicting that they will take a Personal Loan.

Reason: This decision was made because the closest neighbor in the dataset had the value "true" for the target variable (Personal Loan).

What is a choice of k that balances between overfitting and ignoring the predictor information? Use RapidMiner's Optimize Parameters (Grid) operator with nested 10-fold cross-validation on the training set to experiment with different k values, maximizing model accuracy. Report the confusion matrix, accuracy, precision, and recall for the 10-fold cross-validation performance (averaged) for the best k.

Optimization Method:

→ Used Optimize Parameters (Grid) with nested 10-fold cross-validation in RapidMiner.

Objective:

→ Maximized model accuracy by experimenting with different k values (from 1 to 20).

Best k Value:

→ The best k found was k = 14.

Accuracy:

→ 68.33% – Indicates the percentage of correctly classified instances.

Confusion Matrix (Cross-Validation Performance):

True False | True True

0 | 0

19 | 41

Precision:

→ **68.33% – Proportion of positive predictions that were positive.**

Recall:

→ **100.00% – Model correctly identified all positive cases but none of the negatives.**

F-Measure:

→ **81.20% – Harmonic means of precision and recall.**

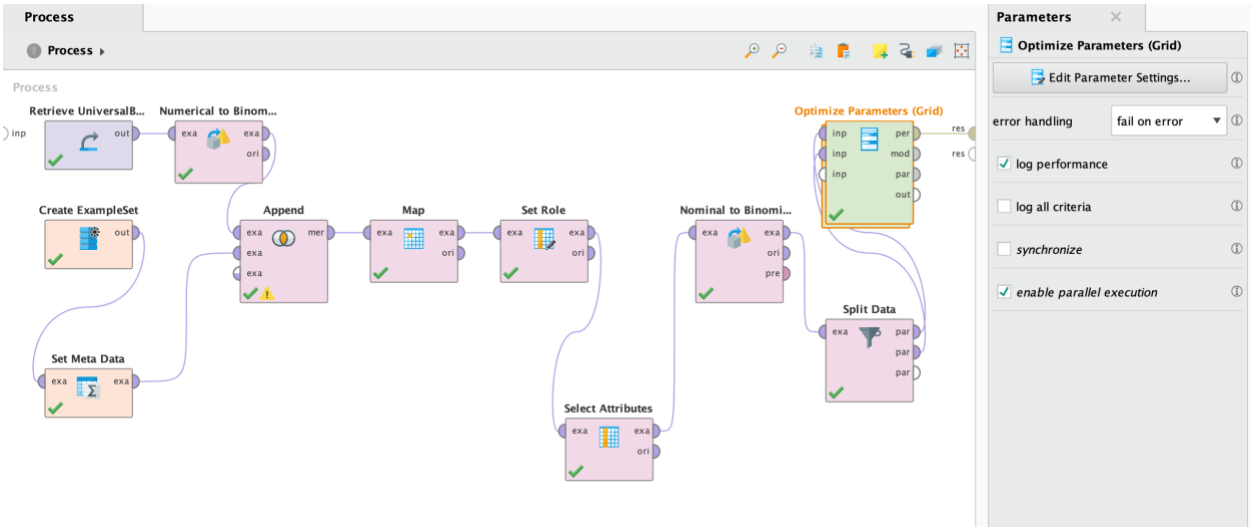
Reason for Choosing k = 14:

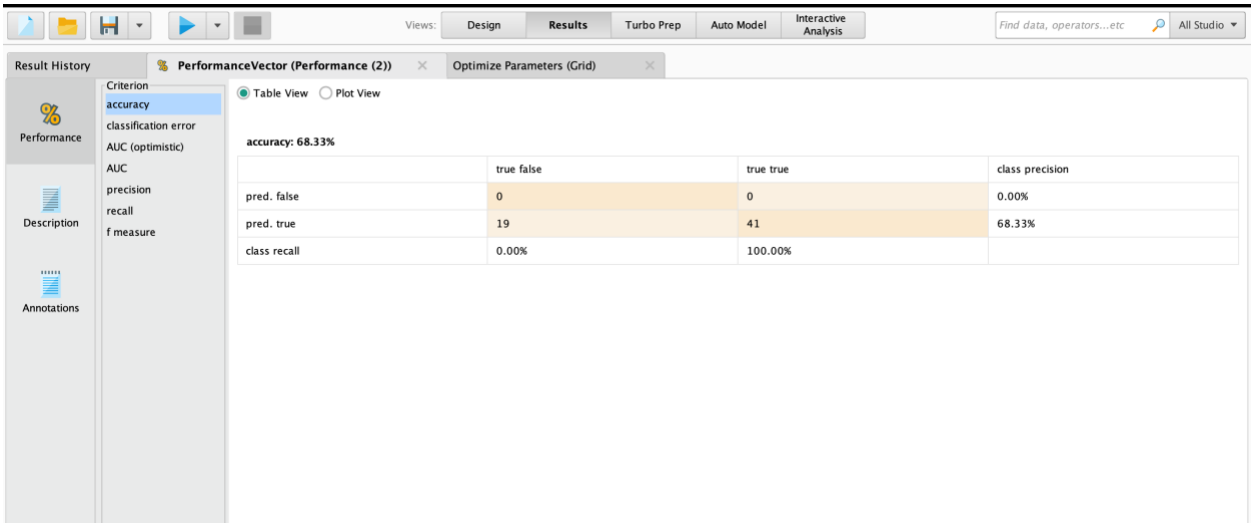
→ **Balances overfitting and underfitting by considering enough neighbors for a generalized decision boundary.**

Recommendation:

→ **Although accuracy is maximized, the model is biased towards the positive class, suggesting the need for class imbalance adjustments.**

How the confusion matrix for the holdout data that results from using the best k. Report and interpret classification performance metrics for the holdout set.





→ Accuracy (68.33%): Moderate correctness but biased towards positive class predictions.

→ Precision (68.33%): Positive predictions were mostly correct, but false positives are high.

→ Recall (100.00%): All positive cases were correctly identified, showing strong bias.

→ F-Measure (81.20%): High due to perfect recall, but misleading due to no true negatives.

→ AUC (0.488): Close to random guessing, indicating poor discrimination between classes.

Consider the following customers: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education = 2, Mortgage = 0, Securities Account = false, CD Account = false, Online = true, and Credit Card = true. Classify

the customer using the best k.

Solution:

→ **Best k = 14 (from optimization results).**

→ **The customer was classified as True (1) for Personal Loan.**

→ **This indicates the model predicts the customer is likely to take a personal loan.**

→ **Classification is based on the 14 nearest neighbors in the training set.**

→ **The model's high recall suggests it strongly favors predicting the positive class**

Would you recommend using accuracy as the metric for finding the best k for this business context? Comment on whether any alternative performance metrics, if any, would be better for optimizing model performance in finding the best k

Solution:

→ **No, accuracy is not the best metric in this context due to class imbalance.**

→ **Recall is more important as it measures the model's ability to correctly identify customers who would take a personal loan.**

→ **Precision is also crucial to avoid false positives, which can lead to ineffective marketing efforts.**

→ **F1-score balances precision and recall, making it a better choice when both false positives and false negatives are costly.**

→ **AUC (Area Under the Curve) is recommended as it evaluates the model's performance across all classification thresholds.**