Lab 6
K-nearest Neighbor based Classification

Objective

Use the given data set to build a K-nearest neighbor (KNN) model and use it to predict the class of given cases.

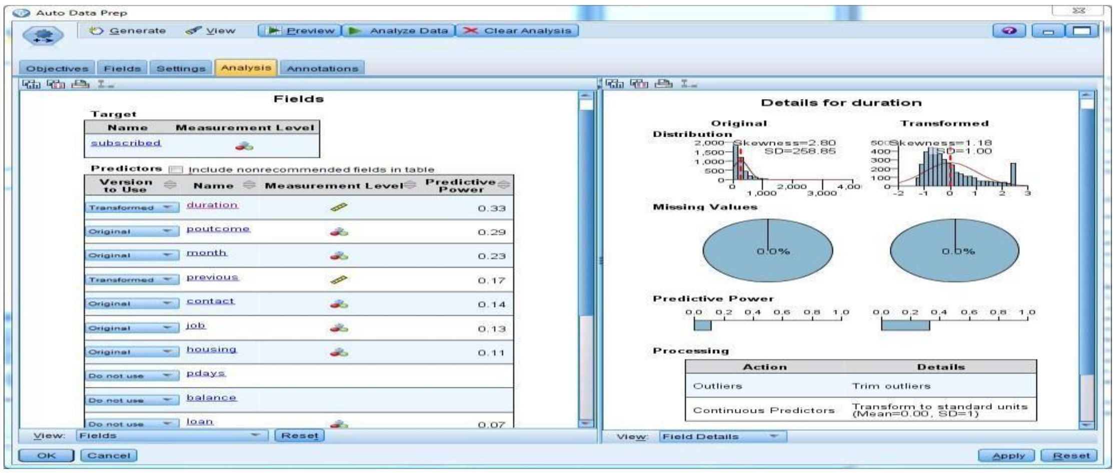
KNN based Classification

The SPSS Modeler offers two type of analysis on KNN model-

- Predict a target field
  - To predict the value of a target field based on the values of its nearest neighbors.
  - User can decide whether speed, accuracy, or a blend of both, are the most important factors when predicting a target field.
- Only identify the nearest neighbors
  - To determine K nearest neighbors for a particular input field

Data Preprocessing

1. Read in the input dataset “bank.csv” using appropriate source node. Select an *"Auto Data Prep"* node from *Field Ops* tab in nodes palette and connect source node to it. Go to *Objectives* tab of the node and select *Optimize for accuracy*. Analyze the *Settings* tab and explore through its options. Go to *Analysis* tab and click on *Analyze Data* button in the menu bar. Select *Fields* option from the drop-down list on left-hand side panel.
2. Select ‘*Do not use’* option from the drop-down list adjoining the field names for fields having *Predictive Power* below a threshold (0.05 in this case). Select *Field Details* on right-hand side panel and explore through details of all the fields one by one.
3. Go to *Generate* menu item in the menu bar of the *Auto Data Prep* node and click on *Filter* node. Connect this node to auto data prep node.



Fields

Target

Name	Measurement Level
subscribed	

Predictors ☐ Include nonrecommended fields in table

Version to Use	Name	Measurement Level	Predictive Power
Transformed	duration		0.33
Original	poutcome		0.29
Original	month		0.23
Transformed	previous		0.17
Original	contact		0.14
Original	job		0.13
Original	housing		0.11
Do not use	pdays		
Do not use	balance		
Do not use	loan		0.07

View: Fields   Reset

Details for duration

Original: Skewness=2.80, SD=258.85
Transformed: Skewness=1.18, SD=1.00

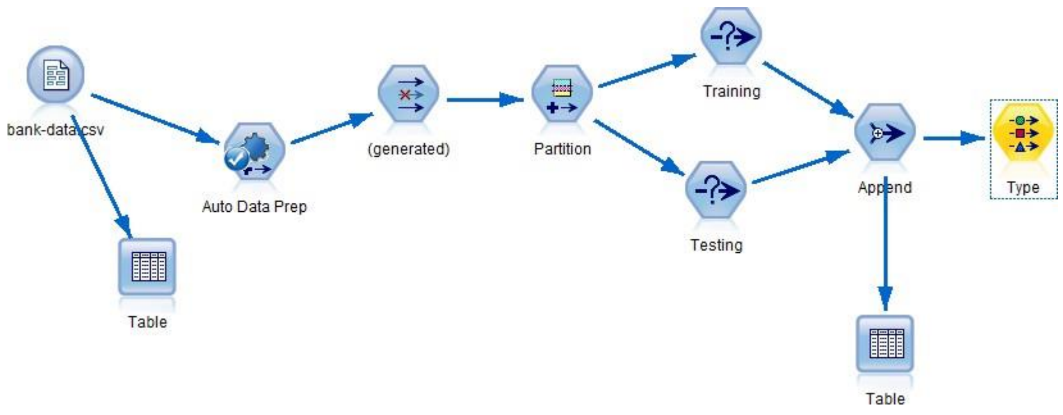
Action	Details
Outliers	Trim outliers
Continuous Predictors	Transform to standard units (Mean=0.00, SD=1)

View: Field Details

Fig 1 - Analysis tab of Auto Data Prep node

Partitioning the dataset:

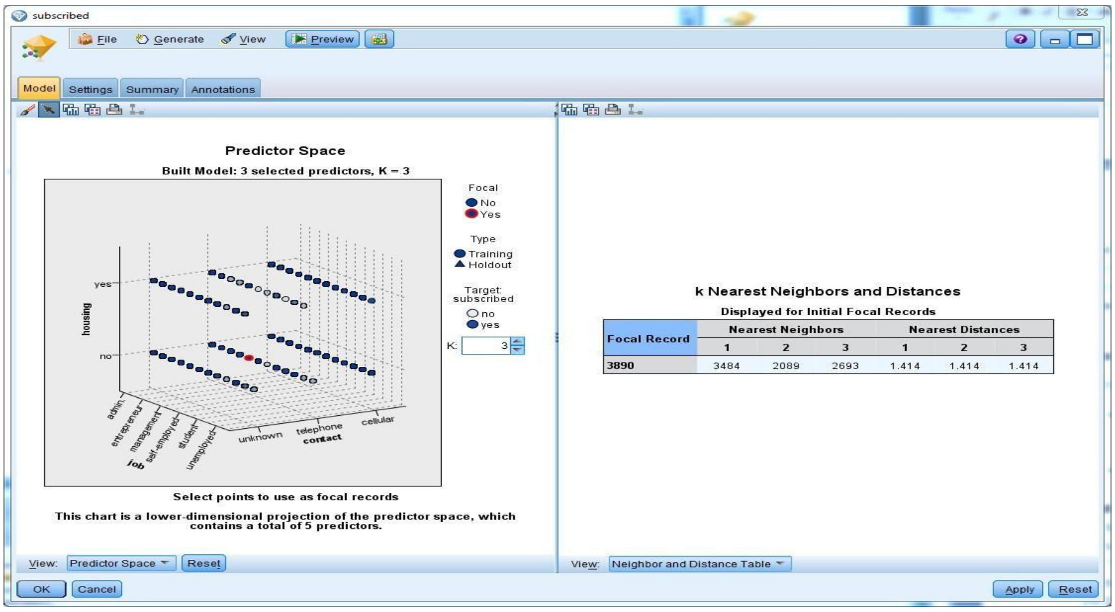
1. Since there is no testing dataset, we need to partition the dataset. In this case, we will partition it in 80:20 ratio using ‘Partition’ node.
2. Select an *"Append"* node from *Record Ops* tab in nodes palette and connect testing and training partitions to it. Go to *"Inputs"* tab of *Append* node and make sure that input dataset is above to be predicted dataset. To view the resulting fields, go to *"Append"* tab.
3. Add a *"Type"* node from *Field Ops* tab in node palette and connect *Append* node to it. Mark the *Role* of target field as *Target* under *Types* tab of this node.



Performing K-nearest neighbor Classification-

1. Select a *"KNN Node"* from *"Classification"* subgrouped under *"Modeling"* tab in nodes palette and connect type node to it Go to *Objectives* tab of this node and select *"Predict a target field"*. Go to *Settings* tab and select *Model* subsection. Uncheck check boxes stating, *"Use partitioned data"* and *"Build model for each split"*. Go to *Neighbors* subsection and specify K=5 and *Euclidean metric* as *Distance Computation* method.

2. Select either perform *Feature Selection* or *V-fold Cross-validation*. Select the later in this case and specify number of folds as 5.*Run* the stream. A *model nugget* with the name of target field will get created.
3. Go to *Model* tab of model nugget. Select any point on the left-hand side panel to see its *Peers Chart* and *Nearest Neighbors and Distance tables* on right-hand side panel.



k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distances		
	1	2	3	1	2	3
3890	3484	2089	2693	1.414	1.414	1.414

Fig 3 - Model tab of Model Nugget

Determining the class of test cases-

1. Connect the Type node to the model nugget to classify the test records.
2. Connect a *Table* node the model nugget and *Run* it. Combined dataset with two additional fields (predicted value of target field and its probability) will be presented.
3. In order to get prediction values of only test cases, connect model nugget to a *Select* node. Go to its *Settings* tab, select *Mode* as *Include* and specify the expression as @PARTITION_FIELD = @TESTING_PARTITION. This will output records which didn't have target-field already predicted i.e. test cases.
4. Since, few fields had been dropped during preparation, in order to get back test cases with all the original fields select a *Merge* node and connect *Select* node of last step and *Source* node of test cases, to it. Go to its *Merge* tab and select *Keys* as *Merge Method* and select all keys as P*ossible keys*.
5. Select a *Table* node and connect *Merge* node to it. Also see the results using Analysis node.

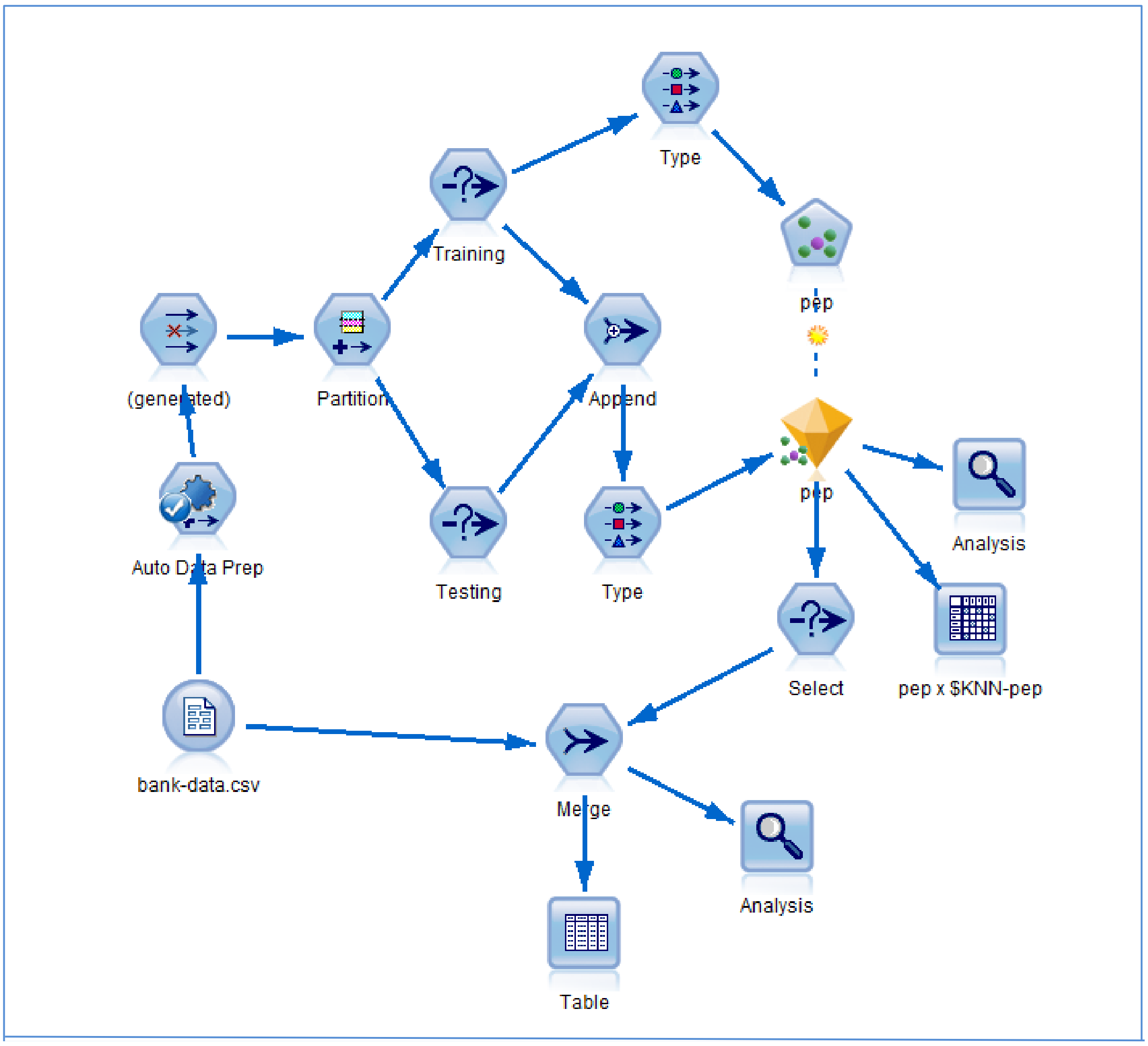


Fig 4 - Complete Stream