PREDICTING PRODUCT SHORTAGE AND SURPLUS

WALMART DATA SET

SHIMEI LIM ABHIJIT SUPREM FANNIE MOK

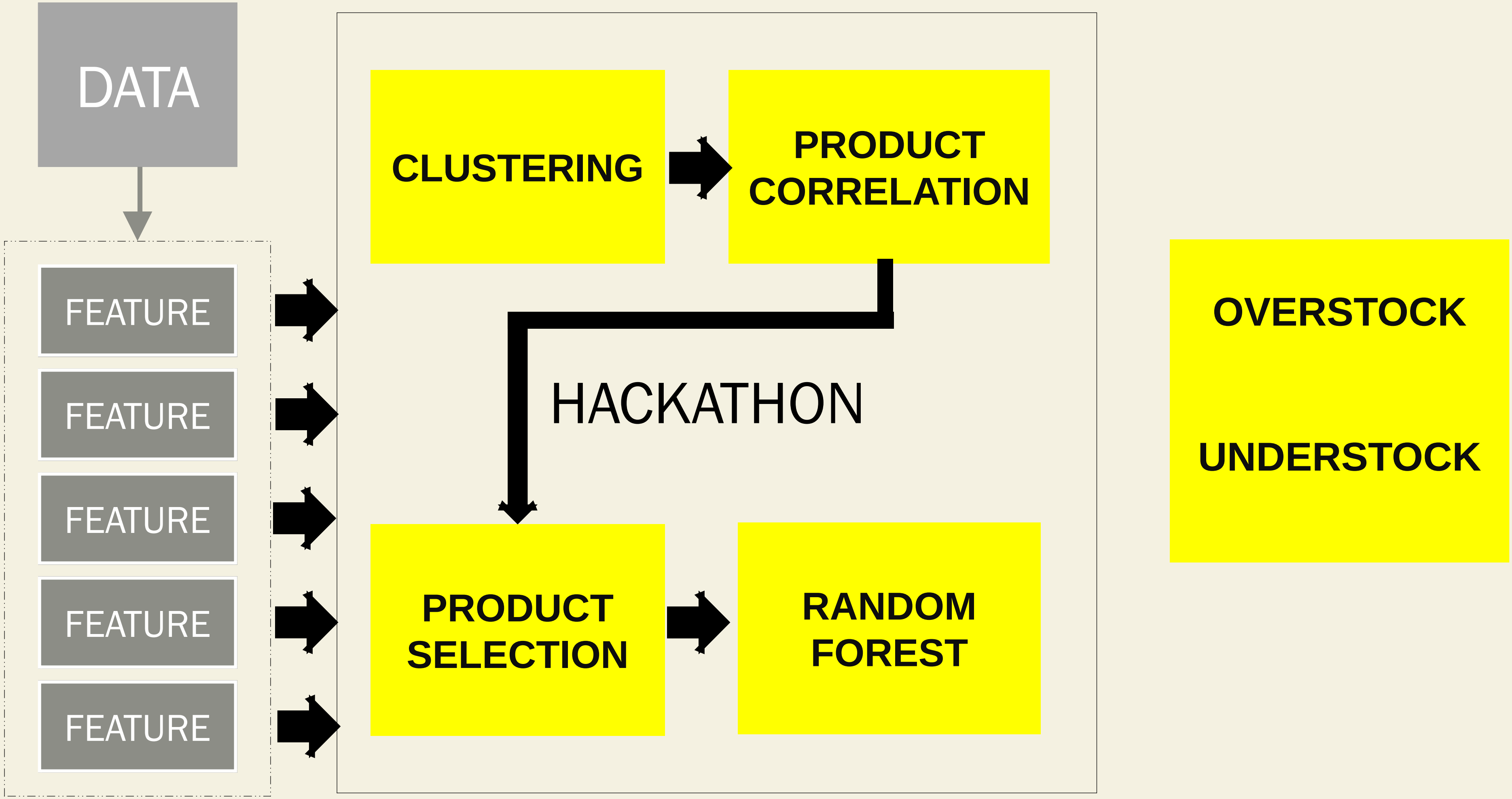
what & why

- Determine product overstock with high accuracy

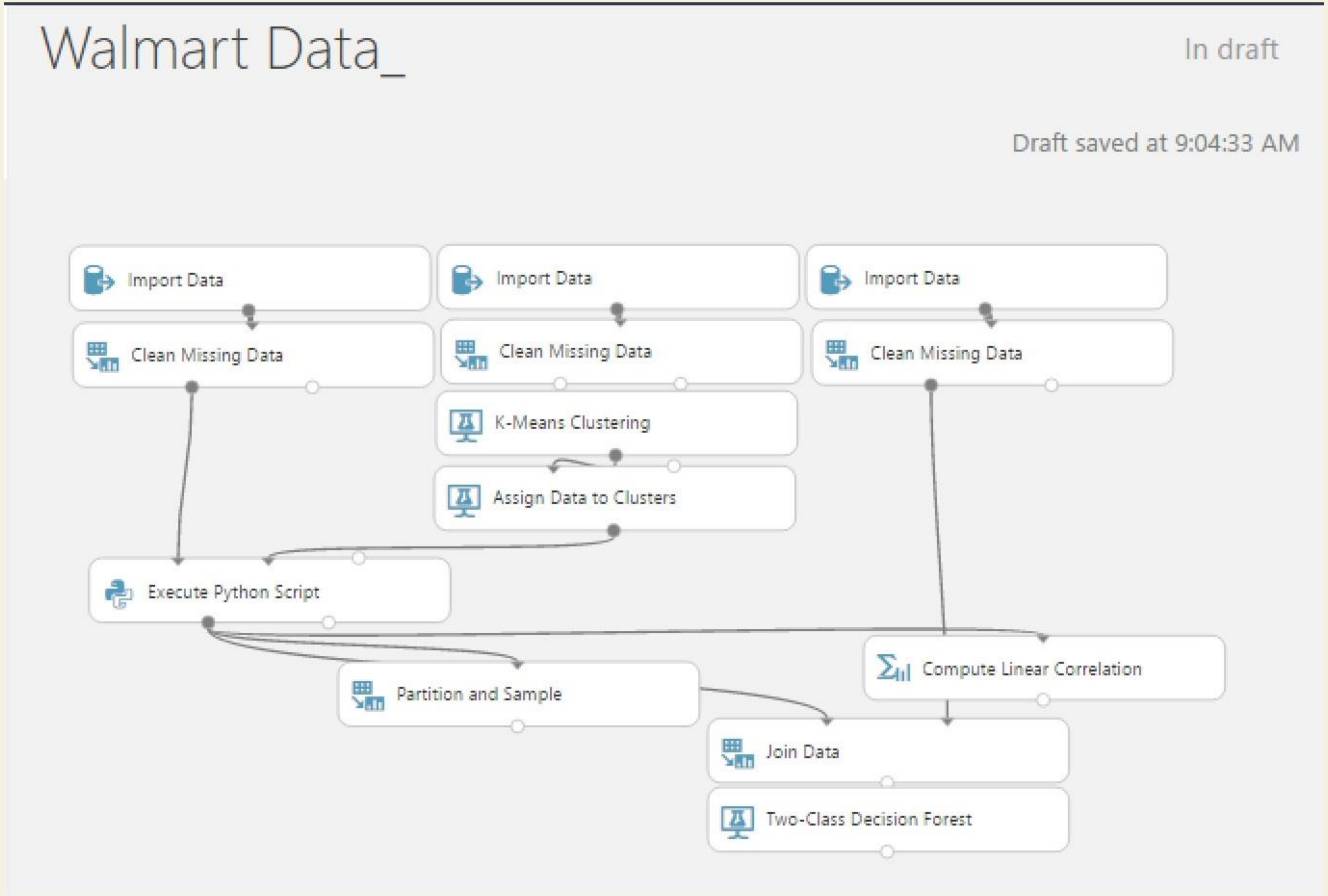
OVERSTOCK	UNDERSTOCK
• Product 12: 98.29%	• Product 19: 97.49%
• Product 19: 96.48%	• Product 60: 99.97%
• Product 45: 95.1%	
• Product 60: 98.6%	

- Relevance: product overstock and understock cause loss in efficiency
- Select relevantly representative products with high correlation to other products
- Products associated with representative store clusters

how


DATA
FEATURE
FEATURE
FEATURE
FEATURE
FEATURE
CLUSTERING
PRODUCT CORRELATION
HACKATHON
PRODUCT SELECTION
RANDOM FOREST
OVERSTOCK
UNDERSTOCK

how



Walmart Data_
In draft
Draft saved at 9:04:33 AM
Import Data
Import Data
Import Data
Clean Missing Data
Clean Missing Data
Clean Missing Data
K-Means Clustering
Assign Data to Clusters
Execute Python Script
Compute Linear Correlation
Partition and Sample
Join Data
Two-Class Decision Forest

where next?

where next?

findings in brief

- Determine product overstock with high accuracy
  - Product 12: 98.29%
  - Product 19: 96.48%
  - Product 45: 95.1%
  - Product 60: 98.6%
  - Product 19: 97.49%
  - Product 60: 99.97%
- Relevance: product overstock an understock cause loss in efficiency
- Select relevantly representative products with high correlation to other products
- Products associated with representative store clusters

procedure

- Data cleanup – delete duplicates, fill missing values
  - *Extraction of relevant data*
- Clustering – Split store types and select representative clusters
  - *Small data set*
  - *Pick clusters with minimum mean and variance of dimension variance*
- Product Correlation – Measure product pair correlations
  - *Pearson correlation of each product pair*
- Product Selection – Correlation peak finding and thresholding
  - *Use absolute correlation*
  - *Correlation threshold to obtain relevantly representative products*
  - *Threshold can be lowered for more prediction options*
- Random Forest -

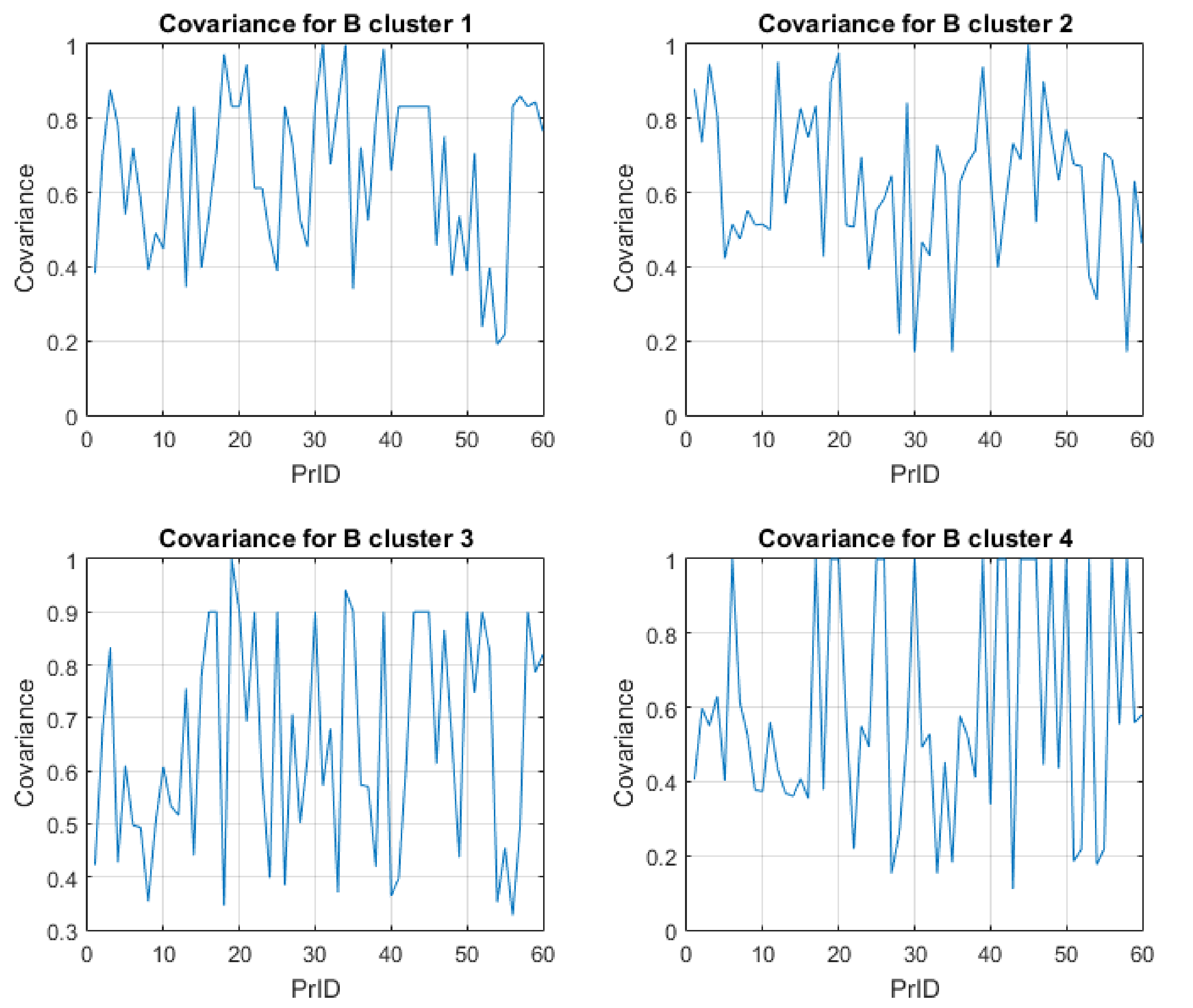
clustering

- Store types: $|A| = 31$; $|B| = 11$; and $|C| = 3$
  - *Ignored type* $|C|$
  - *K-means clustering on A,B separately*
- K-selection
  - *Run multi-trial K-means from K=2 to* $\frac{|type|}{2}$
  - *Obtain dimensional variance of cluster C*
  - *Select clustering with smallest* $\overline{var(C)}$ *and* $var\big(var(C)\big)$
  - *Results vary*

- K-means (labeled clusters in S)
  - $\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \left|\left|x - \mu_i\right|\right|^2$

product correlation

- Store types: $|A| = 31$; $|B| = 11$; and $|C| = 3$
  - *Ignored type $|C|$*
  - *K-means clustering on A,B separately*
- K-selection
  - *Run multi-trial K-means from K=2 to $\frac{|type|}{2}$*
  - *Obtain dimensional variance of cluster C*
  - *Select clustering with smallest $\overline{var(C)}$ and $var\left(var(C)\right)$*
  - *Results vary*

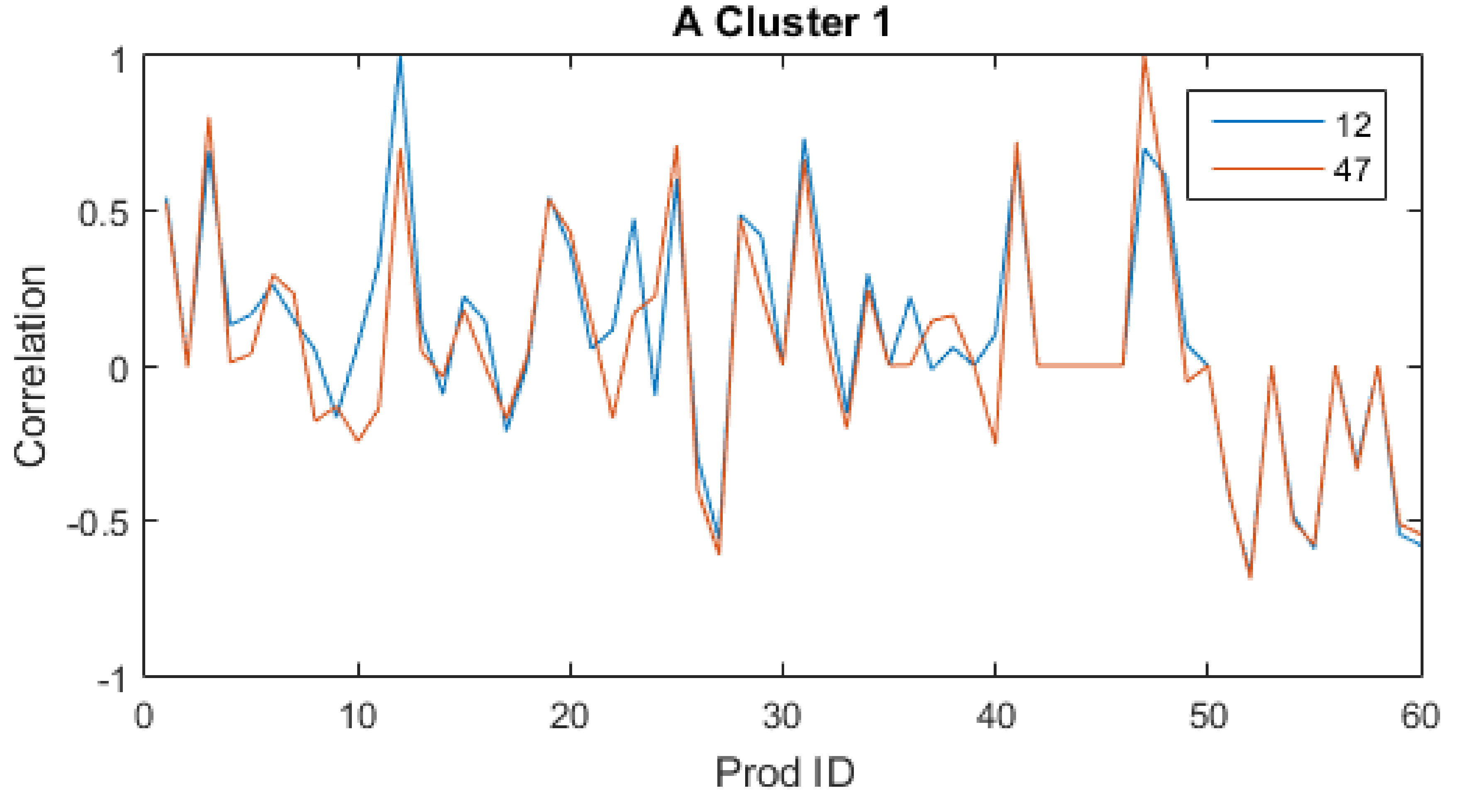


random forest classification

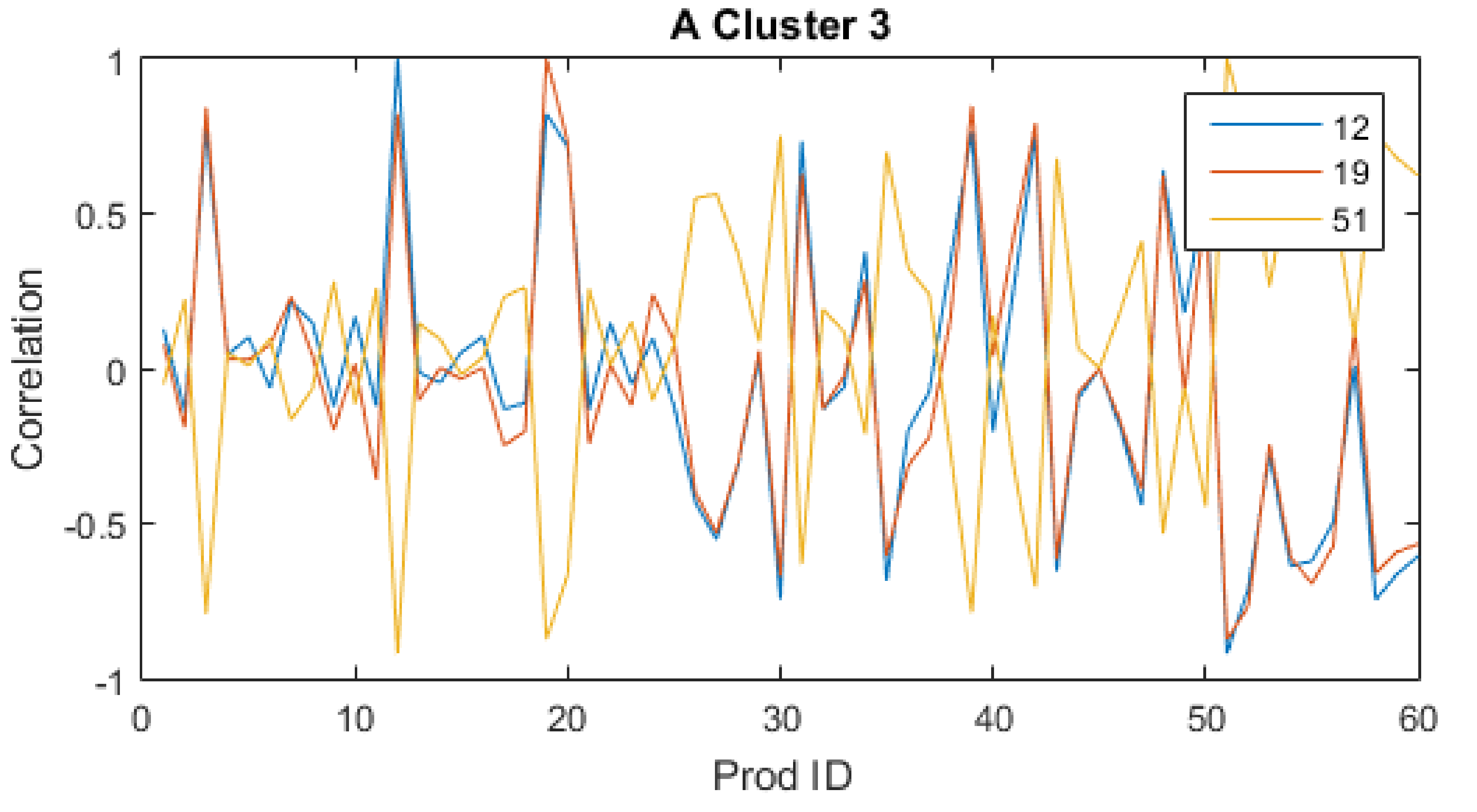
- Ensemble method
  - *Build multiple decision trees to facilitate error-free decisions*
  - *Use Gini impurity index to iterate*
  - $G = \sum_{i=1}^{n_c} p_i(1 - p_i)$
- Does not overfit and maintains higher accuracy versus simpler methods
- Effective in high dimensions with non-binary classification
- Can achieve higher accuracy through robust methods

- Can observe variable interactions to obtain relevant features for dimension reduction

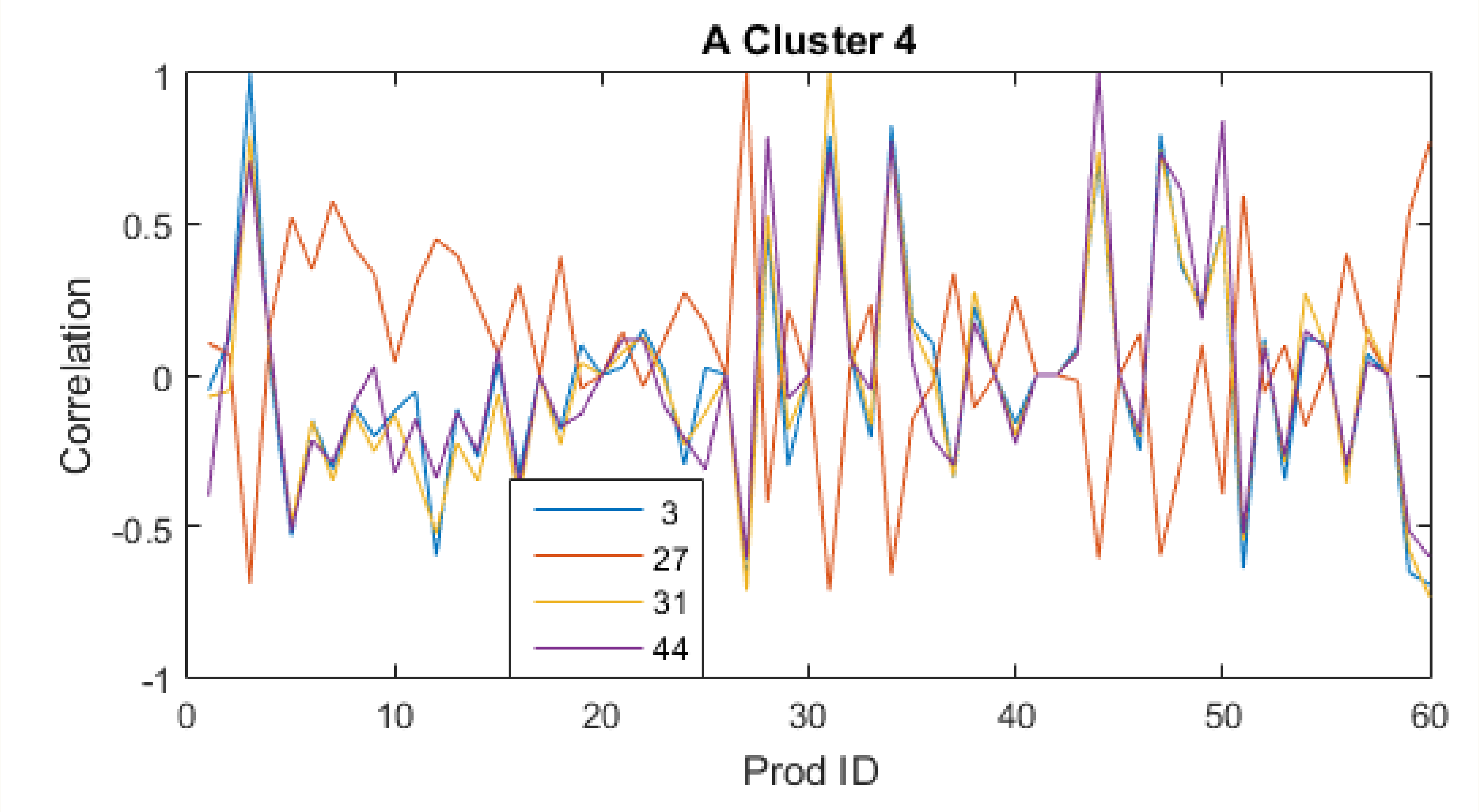
product selection



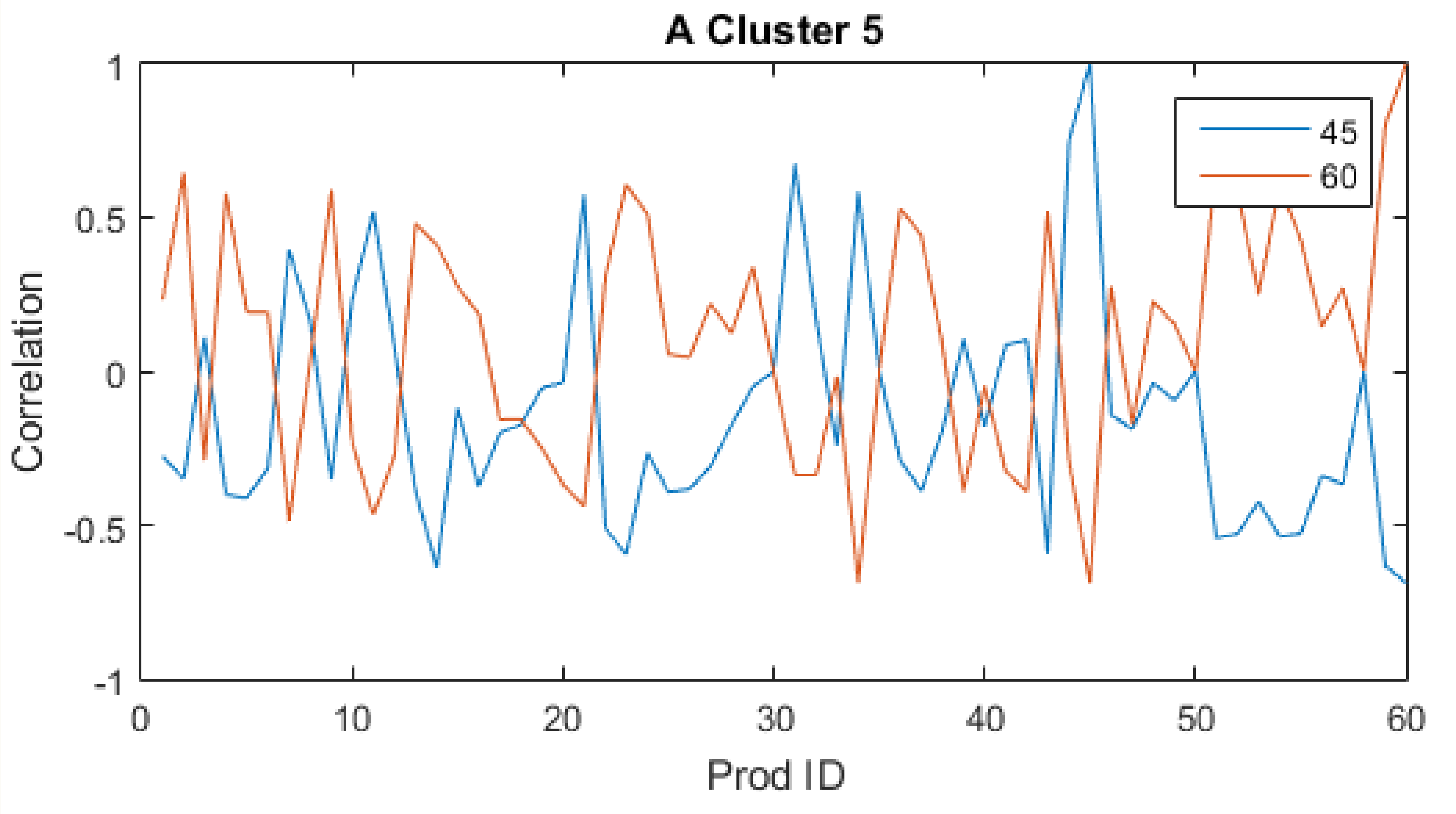
product selection



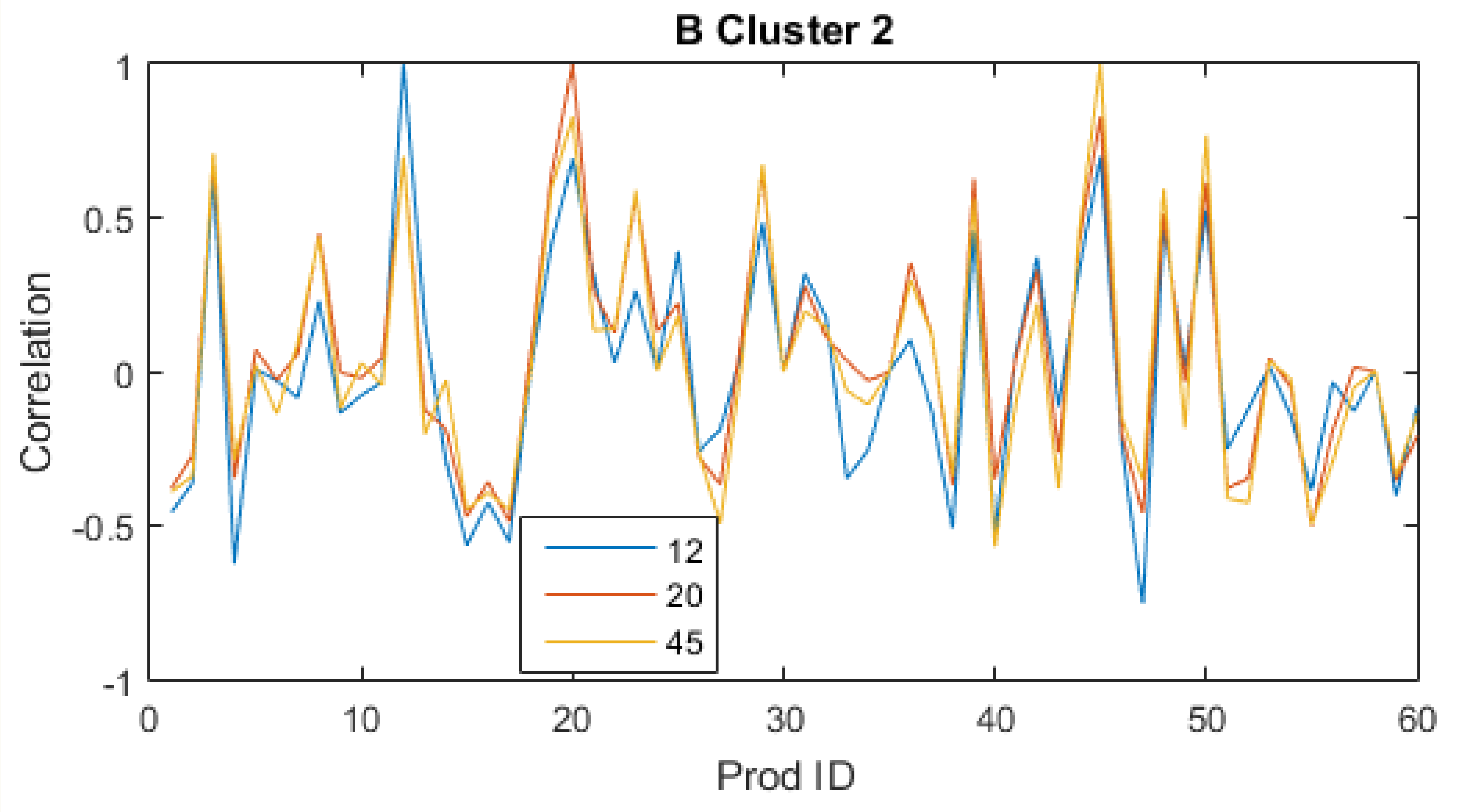
product selection



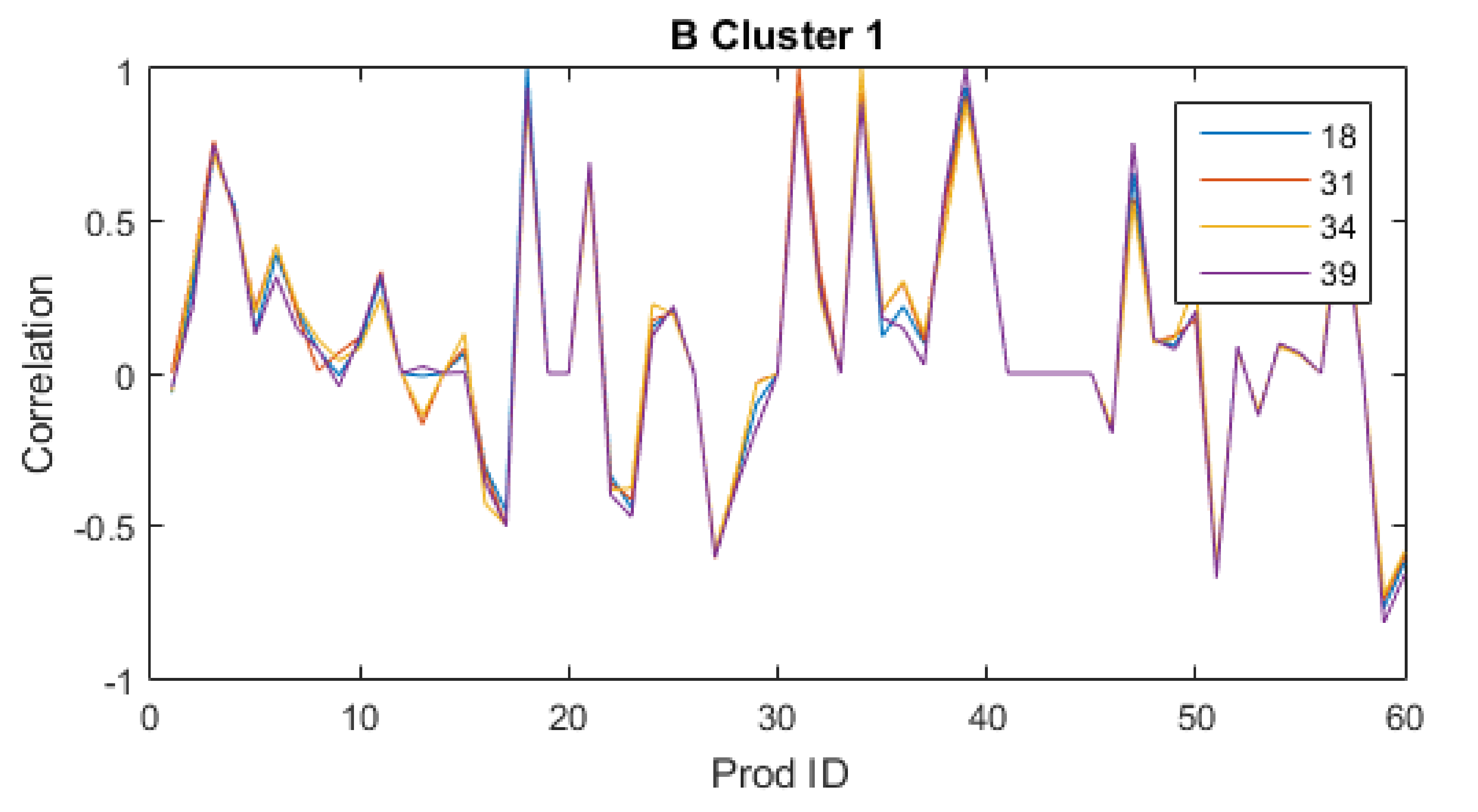
product selection



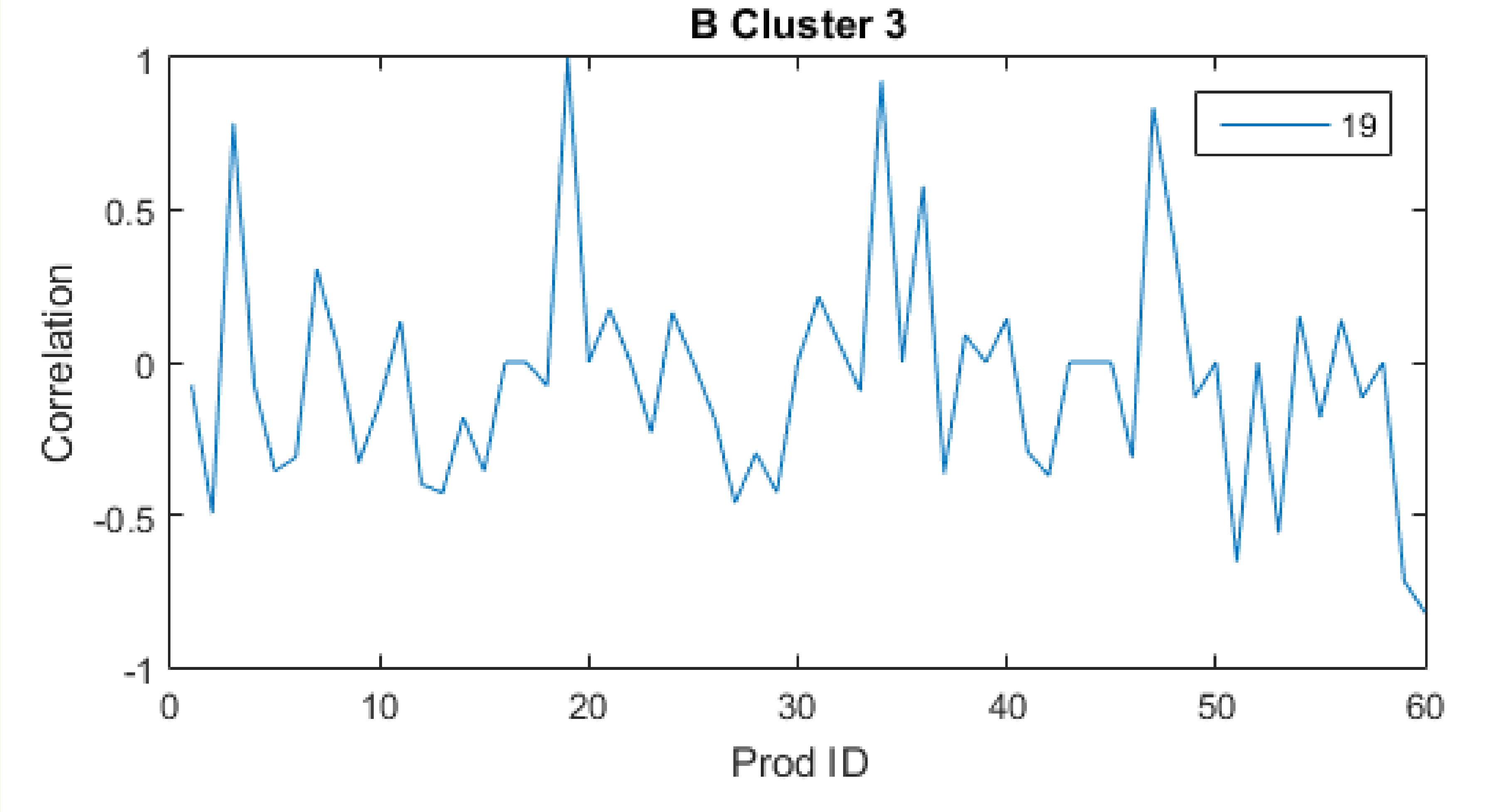
product selection



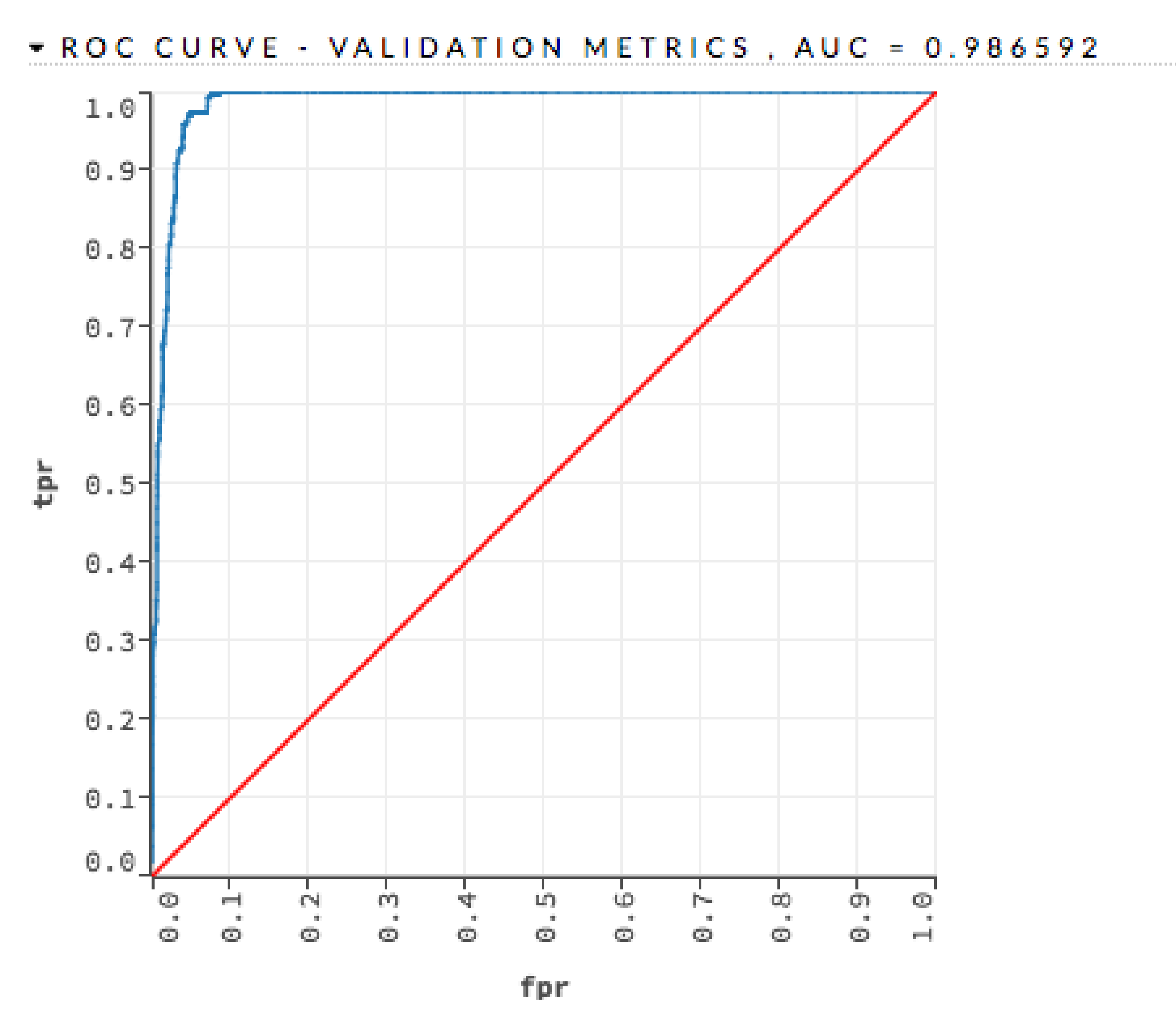
product selection

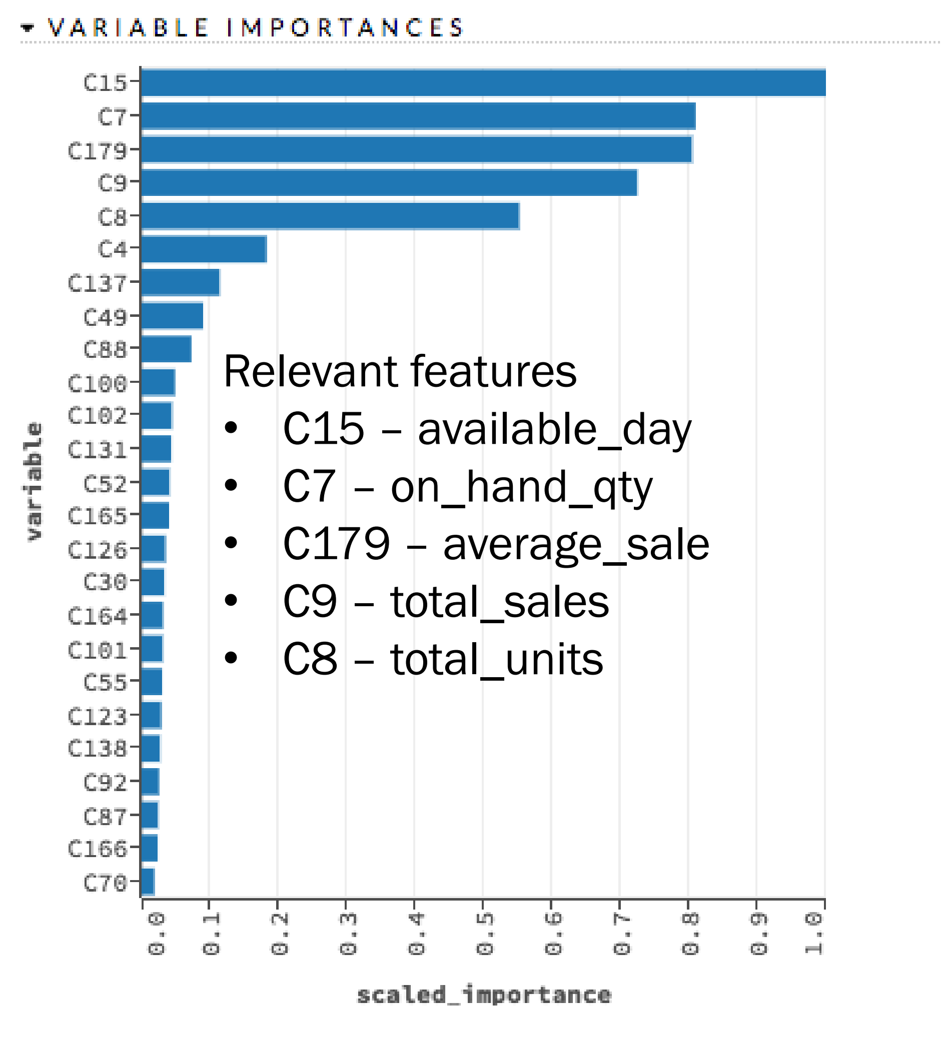


product selection



learning rate – product 60 (RERL)

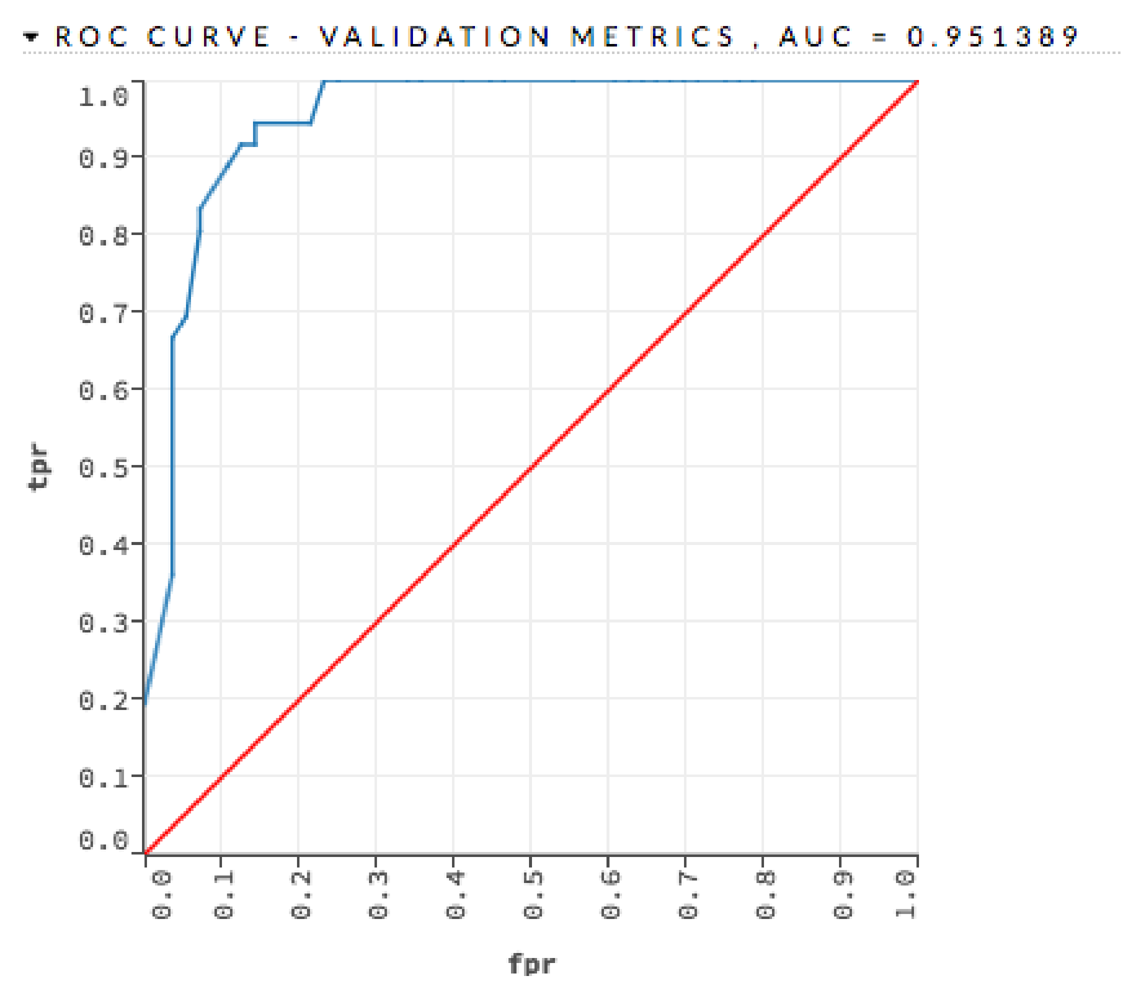


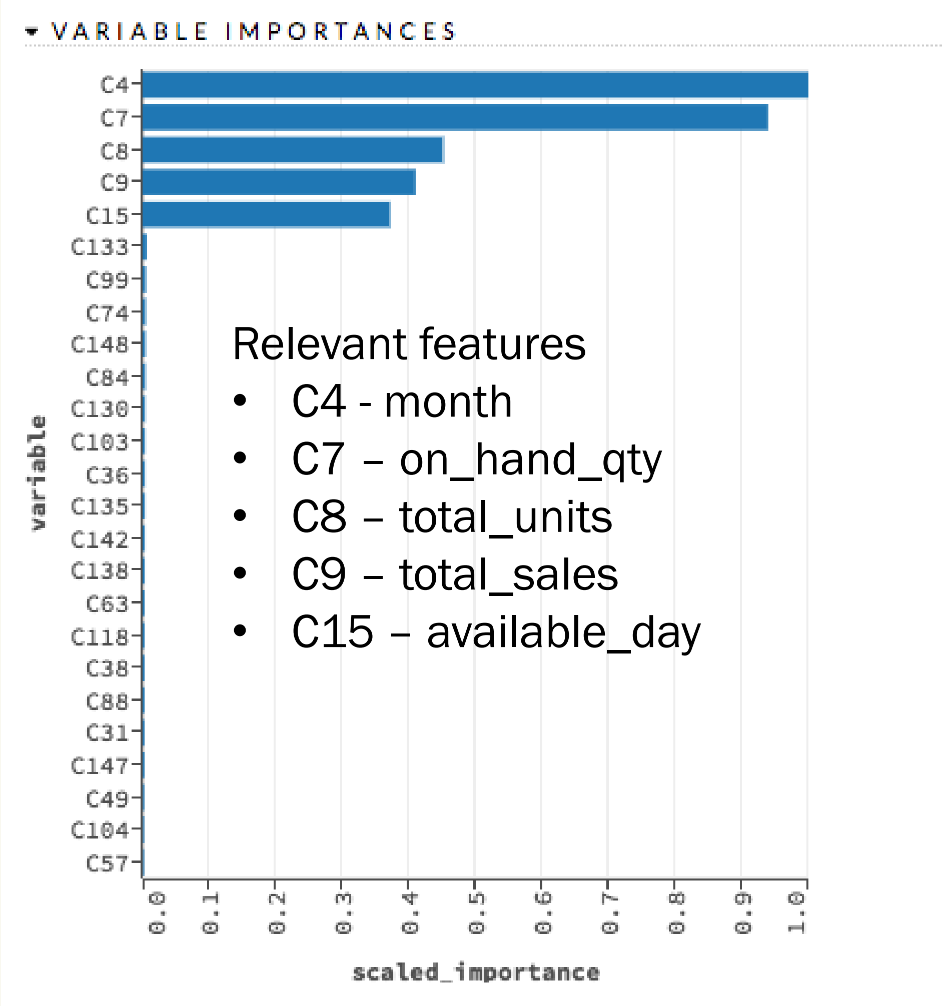


Relevant features

- C15 – available_day
- C7 – on_hand_qty
- C179 – average_sale
- C9 – total_sales
- C8 – total_units

learning rate – product 45 (RERL)

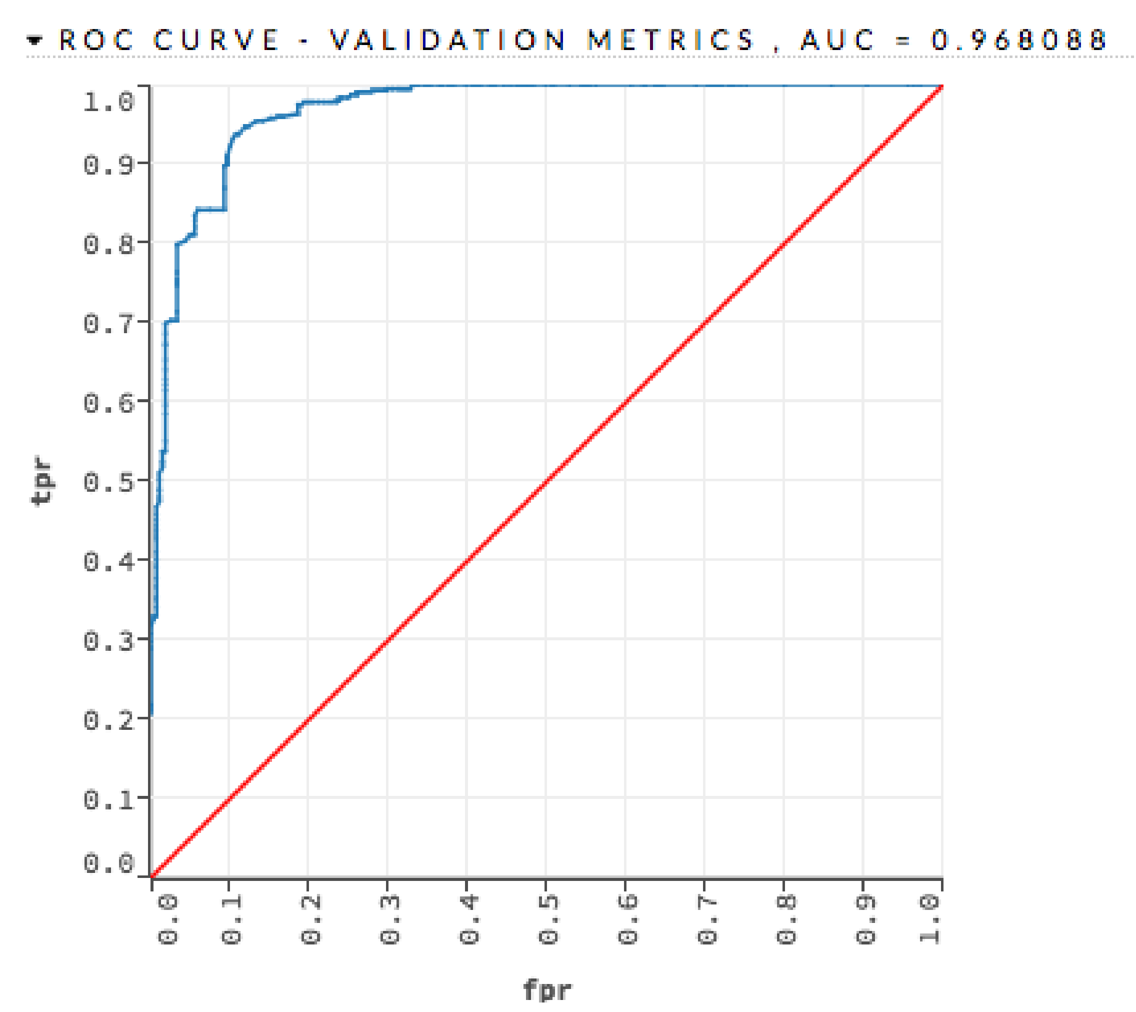


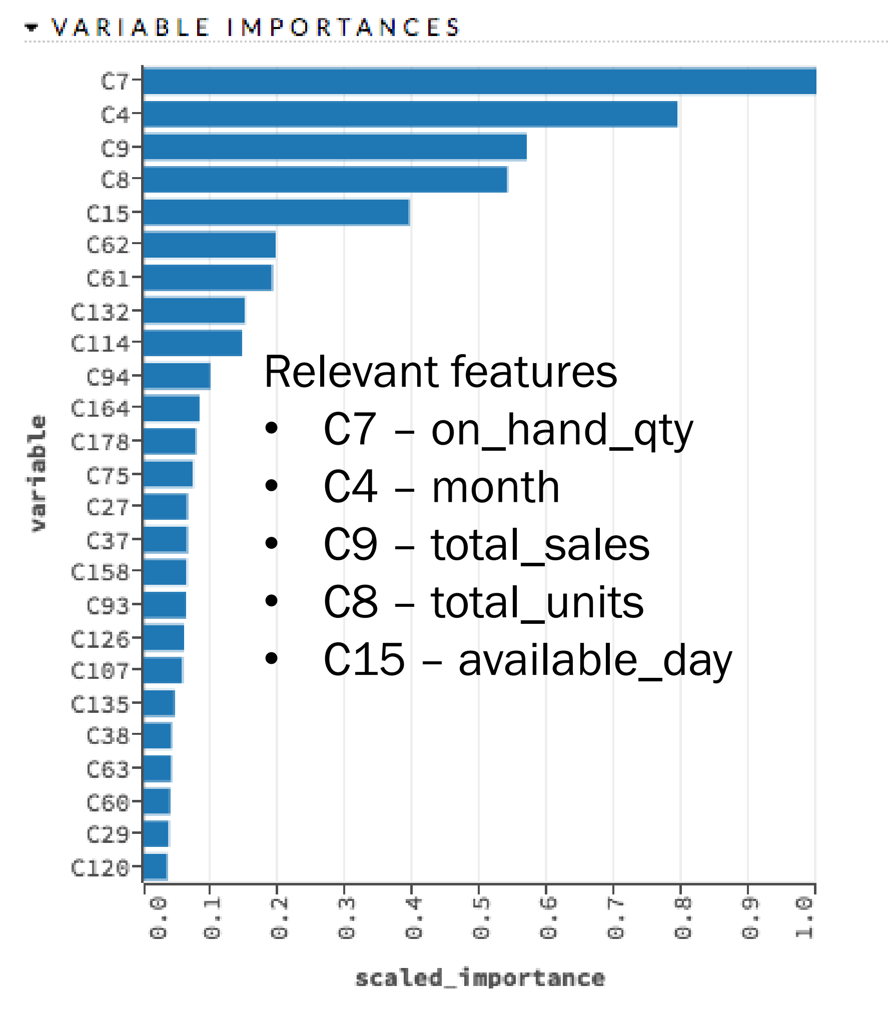


Relevant features

- C4 - month
- C7 – on_hand_qty
- C8 – total_units
- C9 – total_sales
- C15 – available_day

learning rate – product 19 (REPL)

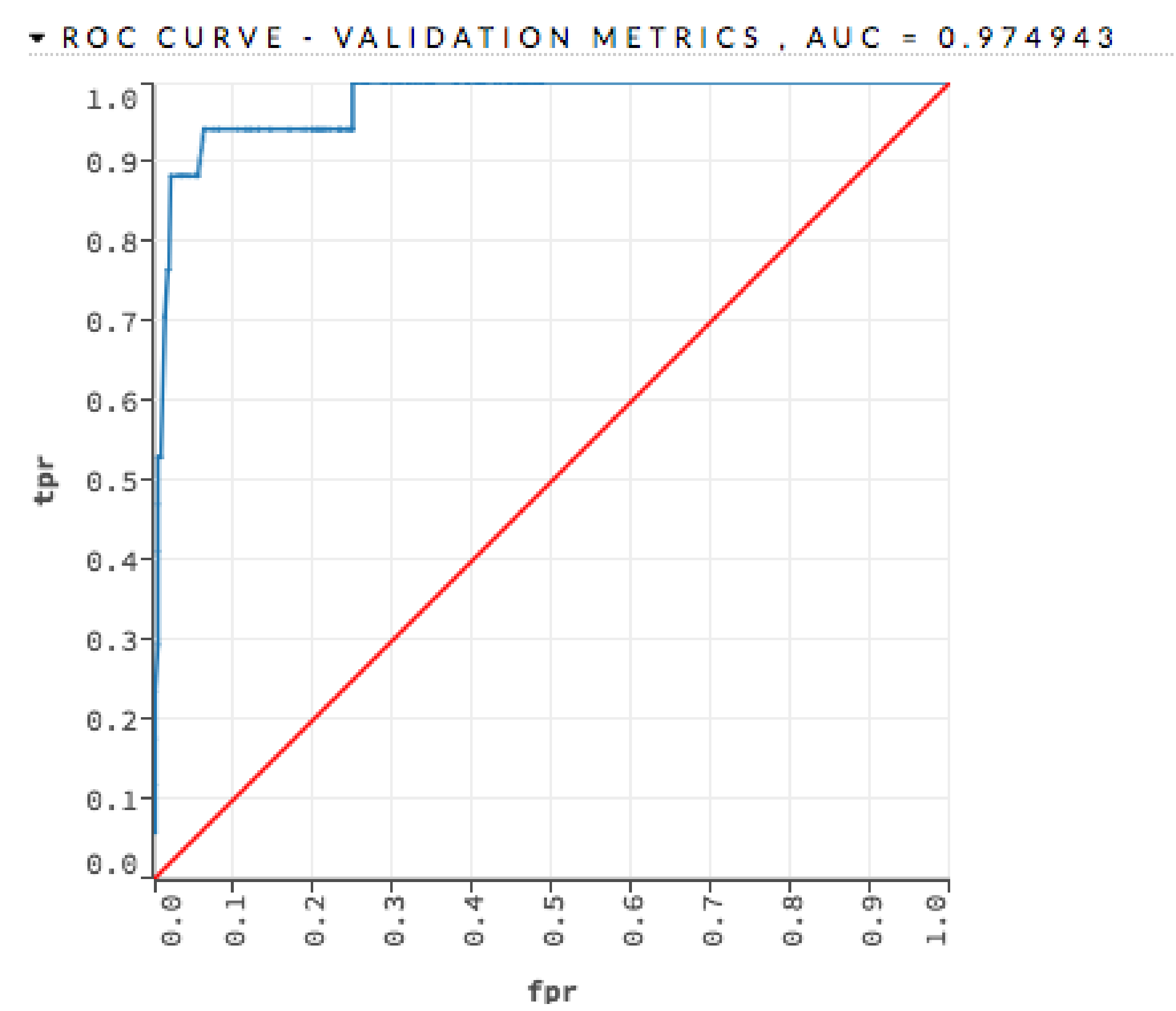


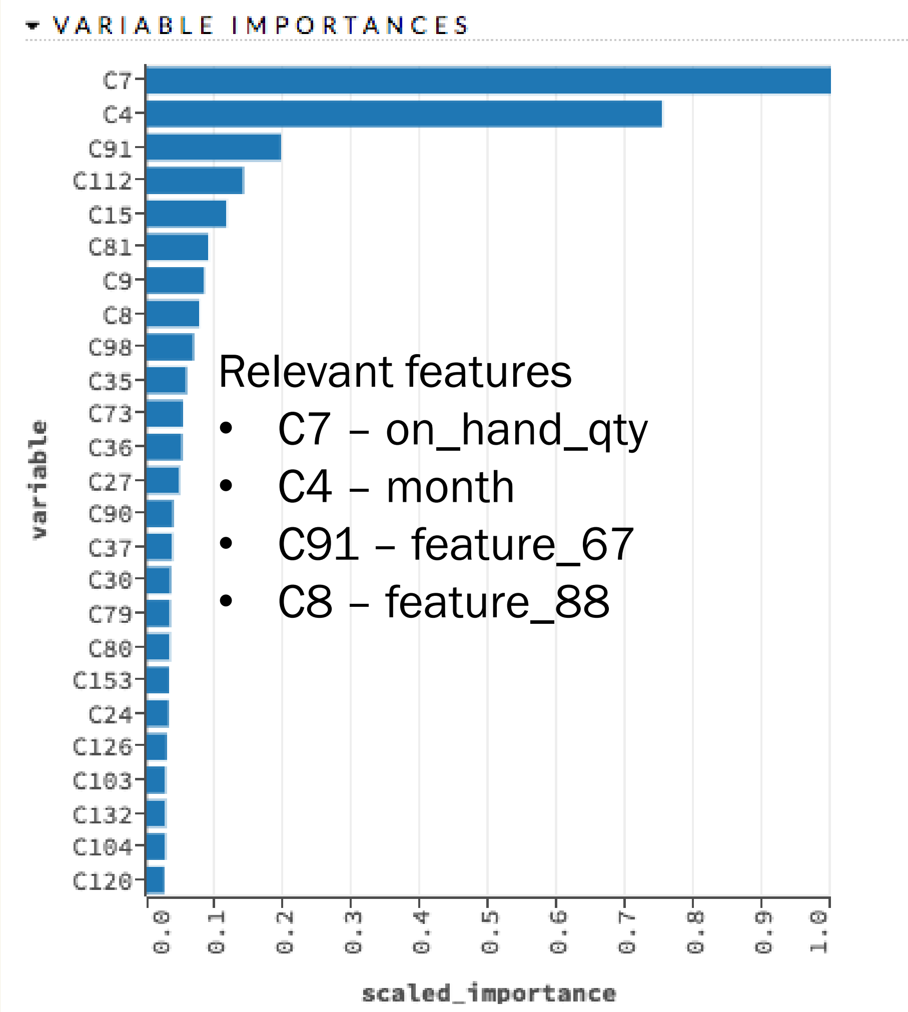


Relevant features

- C7 – on_hand_qty
- C4 – month
- C9 – total_sales
- C8 – total_units
- C15 – available_day

learning rate – product 19 (POS)

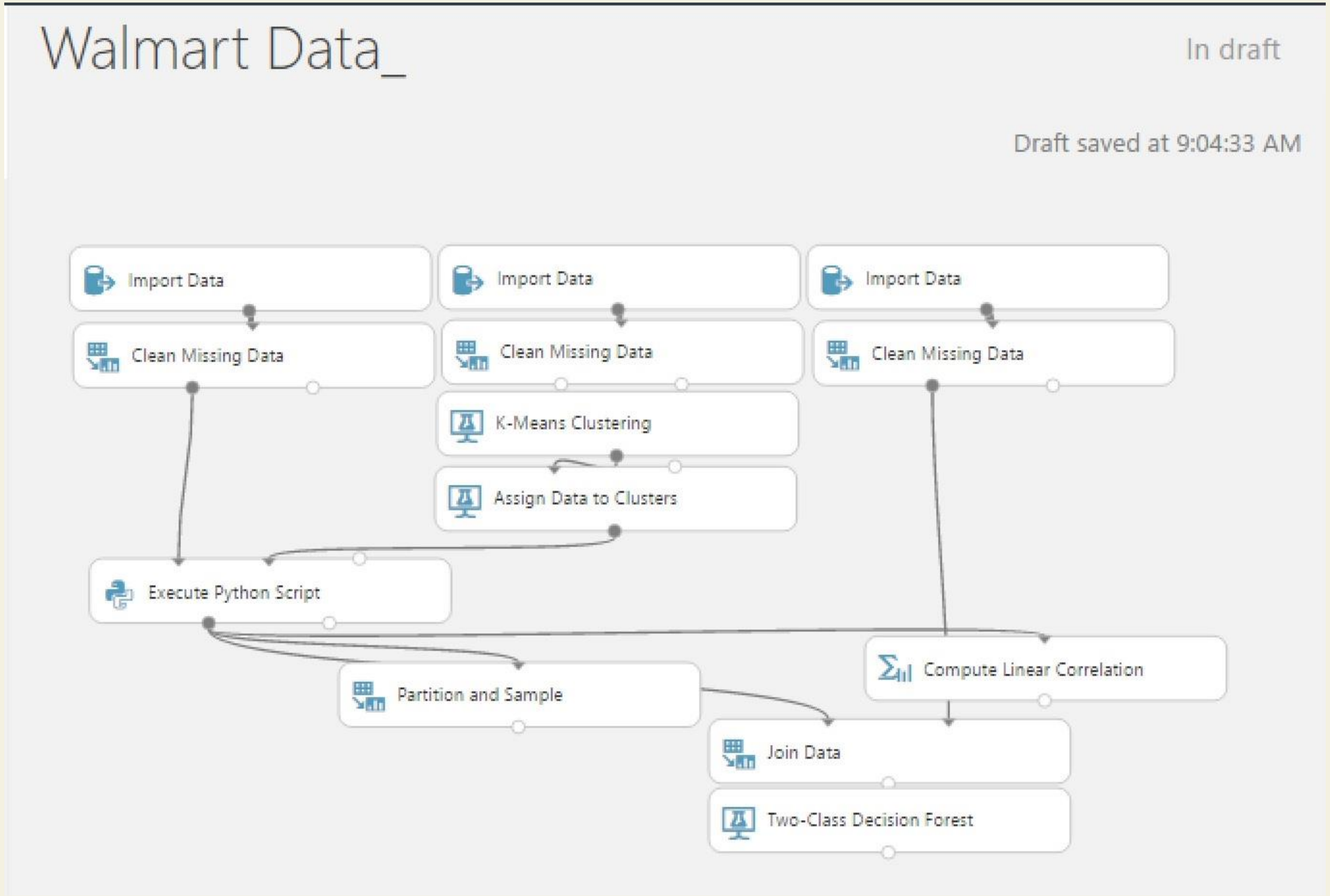




Relevant features

- C7 – on_hand_qty
- C4 – month
- C91 – feature_67
- C8 – feature_88

data flow


Walmart Data_
In draft
Draft saved at 9:04:33 AM
Import Data
Import Data
Import Data
Clean Missing Data
Clean Missing Data
Clean Missing Data
K-Means Clustering
Assign Data to Clusters
Execute Python Script
Compute Linear Correlation
Partition and Sample
Join Data
Two-Class Decision Forest

variance comparisons