Rotman x Scotiabank®

Scotiabank AML Solution

IMI Big Data Competition

Presented by Team 6 Co-MMA

Rotman School of Management
UNIVERSITY OF TORONTO

Alan
Liu

Humza
Butt

Kevin
Zhang

Devashree
Kumar

Team Co-MMA

Master of Management Analytics

Rotman School of Management
UNIVERSITY OF TORONTO

Agenda

1. Background
2. Data Processing and EDA
3. Modeling Approach
4. Clustering & Segmentation
5. Adapt to Big Data
6. Recommendations

1. Background

Background

Rotman

Anti Money Laundering(AML)

Scotiabank®

The Ask:

Which individual customers are considered to be high risk for money laundering?

Background:

- Money laundering is an **illegal action** to use layering of accounts to clean dirty money
- **2 - 5% of Global Economy** are associated with Money laundering
- Scotiabank has labeled a few data with the risk profile for money laundering and wish to use those to **identify all the accounts with high risks**

2. Data Processing & EDA

Preliminary Analysis

Data Preprocessing

Scotiabank® X Rotman

Data Overview

Dataset Used

- Customer data
- Transaction data

Assumption

- Training set is **not a full representation of the big dataset**
  - Emphasis on the **Canadian individual customers**

Preprocessing

1 Data Cleaning

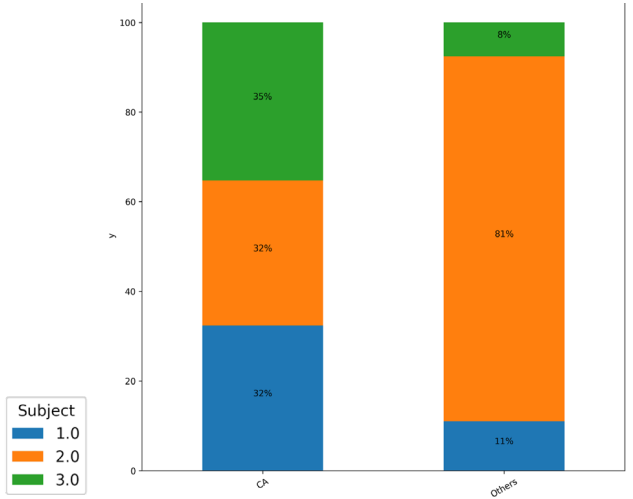
Remove all missing data,
Grouped Categories
- Account Types
- Occupation Status
- Countries

2 Combine Datasets (One-to-One)

Combined with aggregation:
- Count
- Average
- Max
- Min
- Standard Deviation
- Usage count for each type of transaction

3 Data Processing

- Dummy Encoding
- Min-Max Normalized



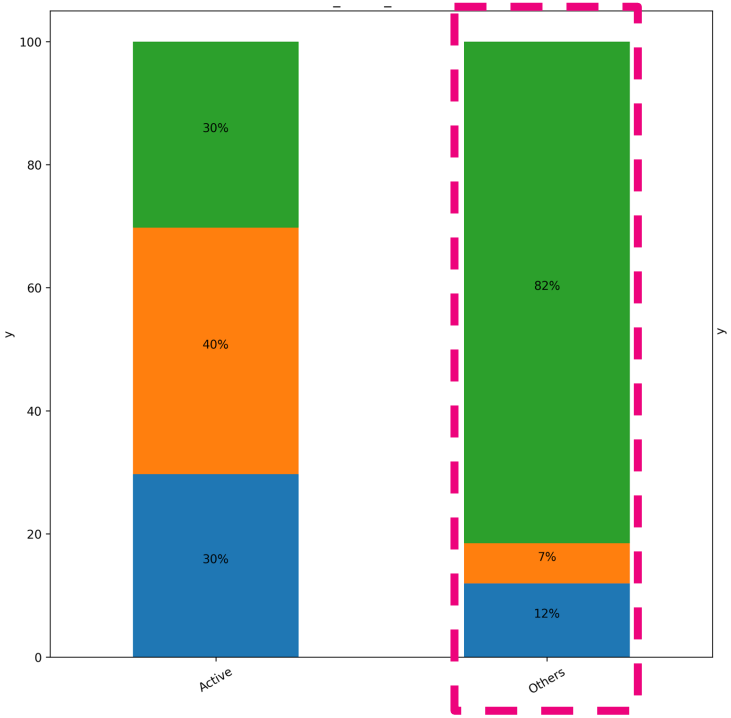
Country: Canada vs Others

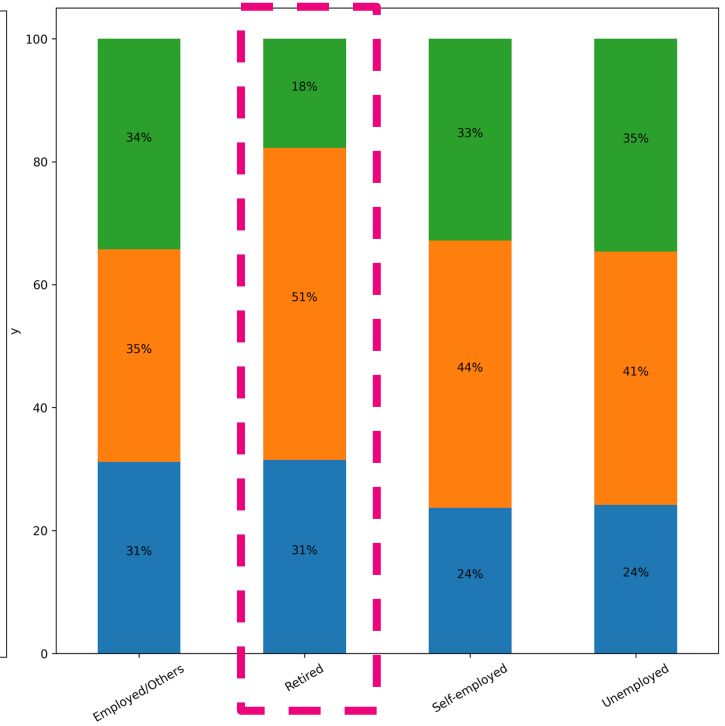
Appendix Attached

Exploratory Data Analysis

Scotiabank® X Rotman

Key factors





Subject
- 1.0 - Low Risk
- 2.0 - Medium Risk
- 3.0 - High Risk

Customer Status:

Active vs Others(Pending, Not a customer & Inactive)

Occupation Code:

Employed, Retired, self-employed, unemployed

Non Active User:

- **high risk** of potential money laundry

Retired

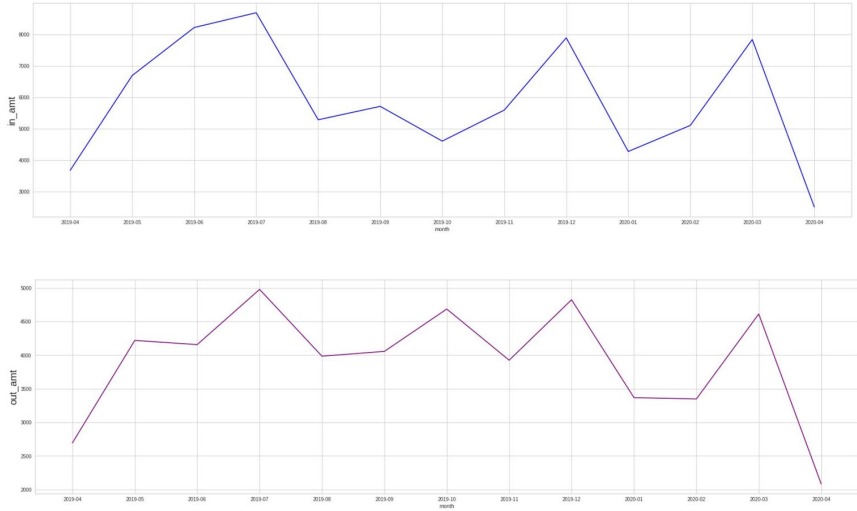
- less likely to have high risk, but have **high medium risk**

Exploratory Data Analysis

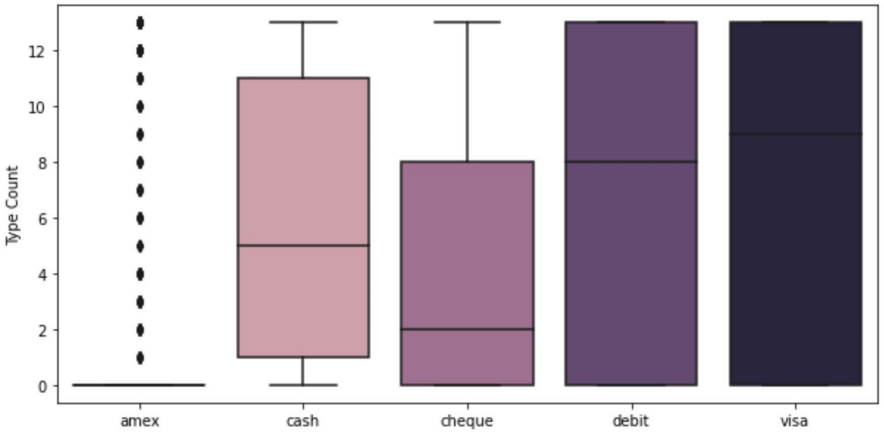
Scotiabank® X Rotman

Data Overview

- Indicates **seasonality** of transactions
  - High value and volume during December holiday seasons



Average Incoming/Outgoing amount :



Transaction Type At Customer Level:

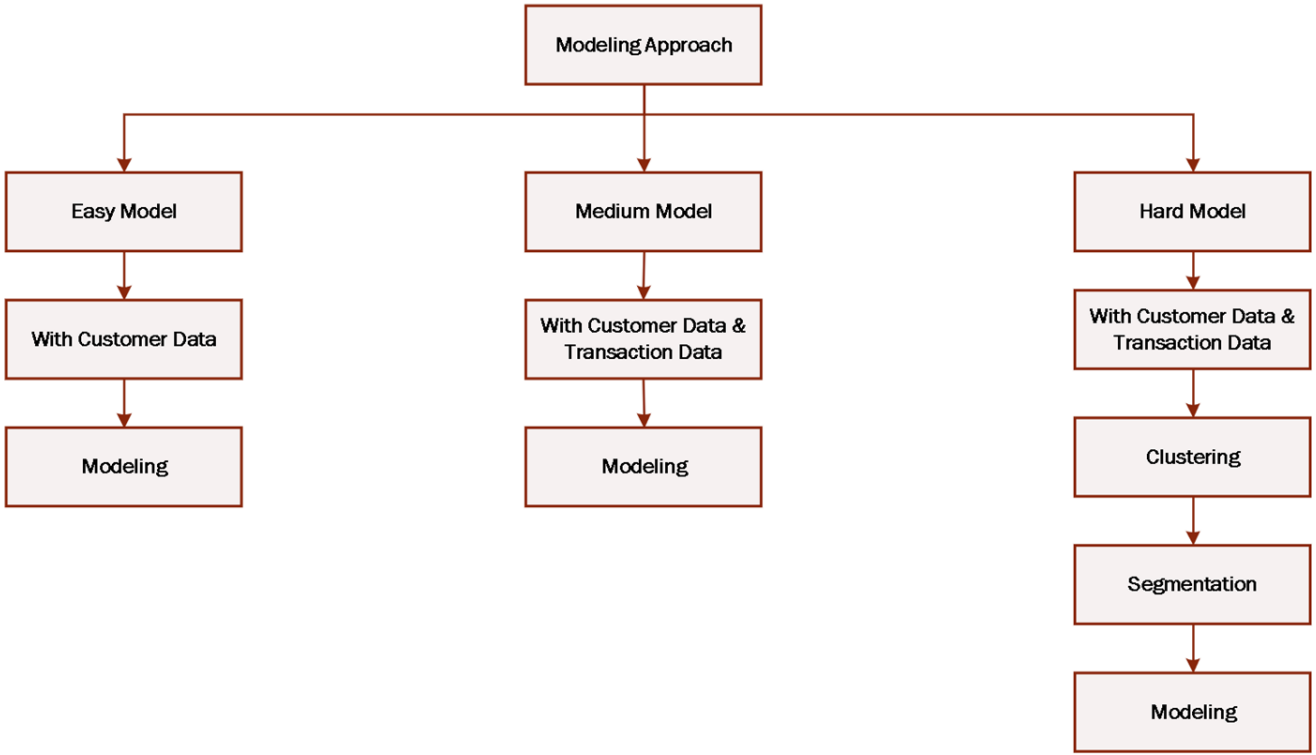
- Individual Canadian customers
  - **Limited use of amex**, as mostly
- Similar statistics for debit and visa

3. Modeling Approach

Easy and Medium Model

Modeling Approach

Approach



Modeling Approach

Scotiabank® X Rotman

Supervised modeling with customer data only (Easy Model)

Model Output by Test Accuracy

Pipeline Models	Training Accuracy	Test Accuracy	Precision	Recall	F-1 score
Logistic Regression	0.429	0.451	0.450	0.450	0.430
Decision Tree	0.431	0.444	0.320	0.440	0.370
KNN	0.413	0.429	0.440	0.430	0.430
SVC	0.434	0.450	0.440	0.450	0.440
Random Forest	0.448	0.456	0.440	0.460	0.440
Gradient Boosting	0.451	0.460	0.450	0.460	0.460

Pipeline Approach

- Pipeline consisted of **6 models for comparison**
- **GridSearchCV** was utilized to find the optimal parameters for each model.
- Best model was determined in terms of **test set accuracy**
- F1 score, Recall, Precision, and AUC metrics were utilized as well for comparison

Best Model Overview

Test Accuracy: 46% **AUC: 65.20%**

Key Models	Relative Accuracy	Reasoning
Gradient Boosting	★★★★★	learn from the residuals
KNN	★	curse of dimensionality
Logistic Regression	★★	multi-label classification

Modeling Approach

Scotiabank® X Rotman

Supervised modeling with customer data only (Easy Model)

Model Overview: Neural Network Model

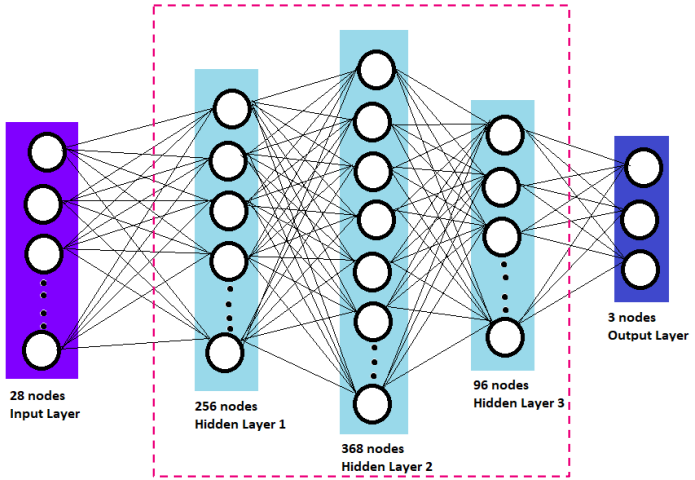
- Neural network was performed with 1 input layer, 3 hidden layers and an output layer.
- Keras Tuner was utilized to search and select the optimal number of neurons.
- The 3 hidden layers has **256, 368 and 96 neurons**.
- The output layer has 3 target variables which is the rating separated into **3 response variables** and the activation function used is 'softmax'**.**
- Test Accuracy : **47.58%**

Hyperparameter Tuning:

- The number of neurons that are used in the hidden layer was chosen between 16 to 512
- Activation function types in the hidden layer were tested using relu and sigmoid
- Model optimizers were tested using rmsprop

Model Output (Neural Network Model)

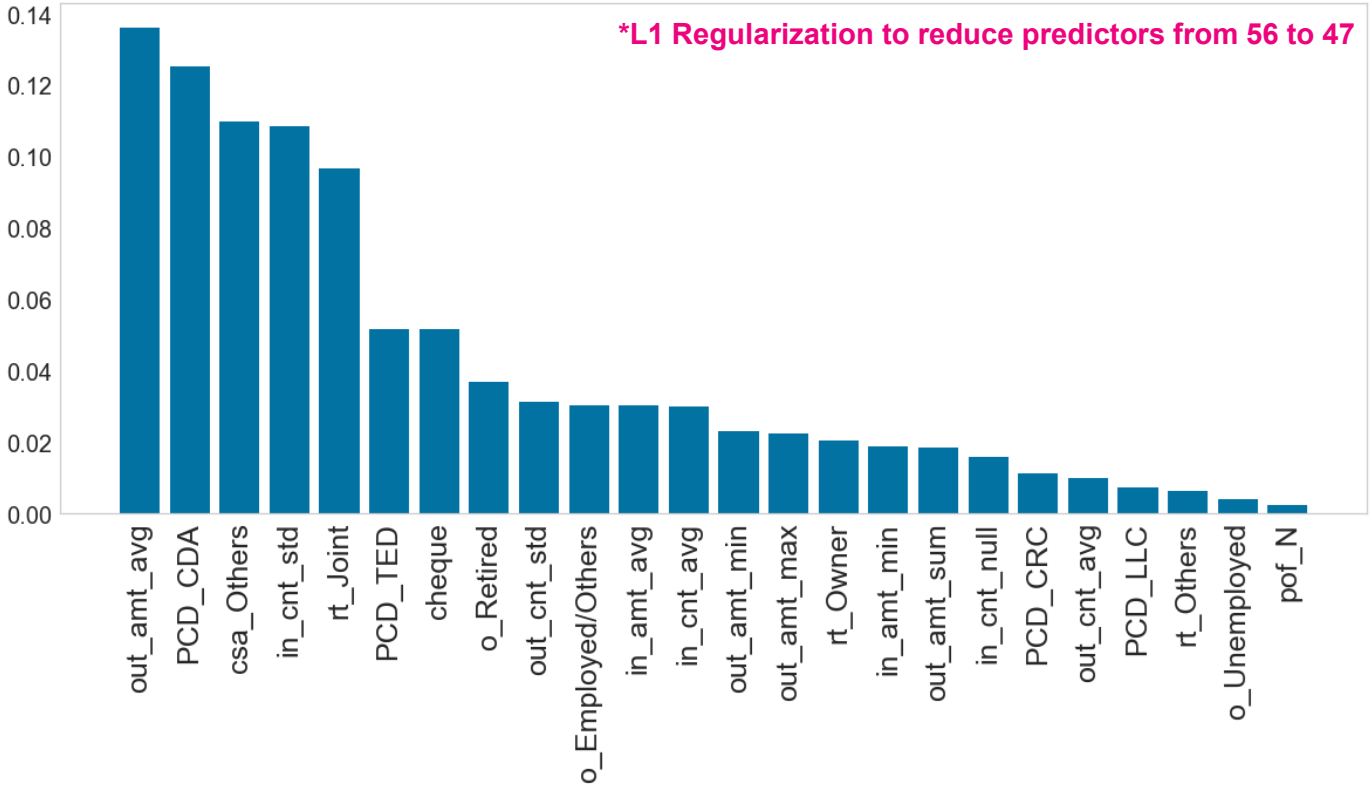
OPERATION		DATA DIMENSIONS	WEIGHTS(N)	WEIGHTS(%)
Input	#####	28		
Dense	XXXXX	----------	7424	5.4%
relu	#####	256		
Dense	XXXXX	----------	94576	68.7%
relu	#####	368		
Dense	XXXXX	----------	35424	25.7%
relu	#####	96		
Dense	XXXXX	----------	291	0.2%
softmax	#####	3		



Modeling Approach

Rotman

Feature Importance



Important Predictors

1. Average Outgoing Amount
2. Count of Chequing Accounts
3. Customer Status: Others
4. Standard Deviation of Incoming Transaction Count
5. Joint Account Type

Modeling Approach

Scotiabank® X Rotman

Supervised modeling with transaction data (Medium Model)

Model Output by Test Accuracy

Pipeline Models	Training Accuracy	Test Accuracy	Precision	Recall	F-1 score
Logistic Regression	0.449	0.452	0.440	0.450	0.440
Decision Tree	0.456	0.438	0.440	0.440	0.440
KNN	0.431	0.434	0.430	0.430	0.420
SVM	0.461	0.457	0.450	0.460	0.450
Random Forest	0.480	0.444	0.430	0.440	0.420
Gradient Boosting	0.456	0.456	0.450	0.460	0.450

Best Model Overview

- 70% of the dataset was used to train the model and 30% of the dataset for testing the model.
- **Test Accuracy:** 46%
- **AUC:** 64.60%
- **PARAMETERS**
  - Learning Rate: **0.1**
  - Loss function: **deviance**
  - Number of estimators: **100**

Modeling Approach

Scotiabank® X Rotman

Supervised modeling with transaction data (Medium Model)

Model Overview: Neural Network Model

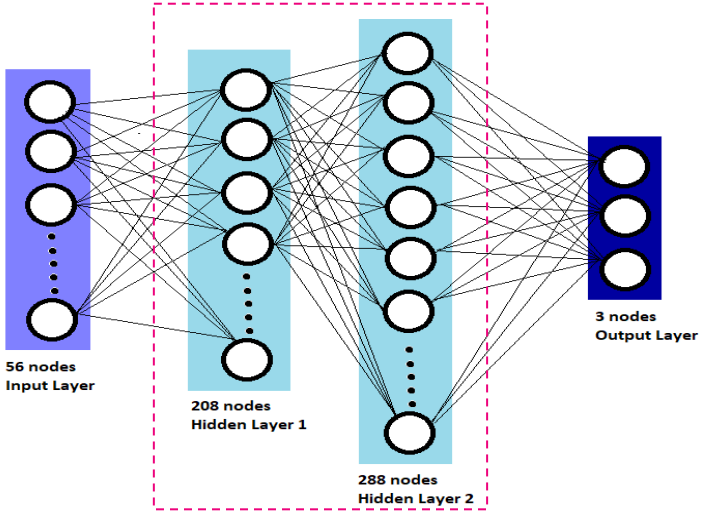
- Neural network was performed with 1 input layer, 2 hidden layer and an output layer.
- Keras Tuner was utilized to search and select the optimal number of neurons.
- The 2 hidden layer has **208 and 288 neurons**.
- The output layer has 3 neurons for each of the 3 ratings of 1,2 and 3.
- Accuracy : **48.59%**

Hyperparameter Tuning:

- The number of neurons that are used in the hidden layer was chosen between 16 to 512
- Activation function types in the hidden layer were tested using relu and sigmoid
- Model optimizers were tested using rmsprop

Model Output

OPERATION		DATA DIMENSIONS	WEIGHTS(N)	WEIGHTS(%)
Input	#####	56		
Dense	XXXXX	----------	11856	16.3%
sigmoid	#####	208		
Dense	XXXXX	----------	60192	82.6%
sigmoid	#####	288		
Dense	XXXXX	----------	867	1.2%
softmax	#####	3		



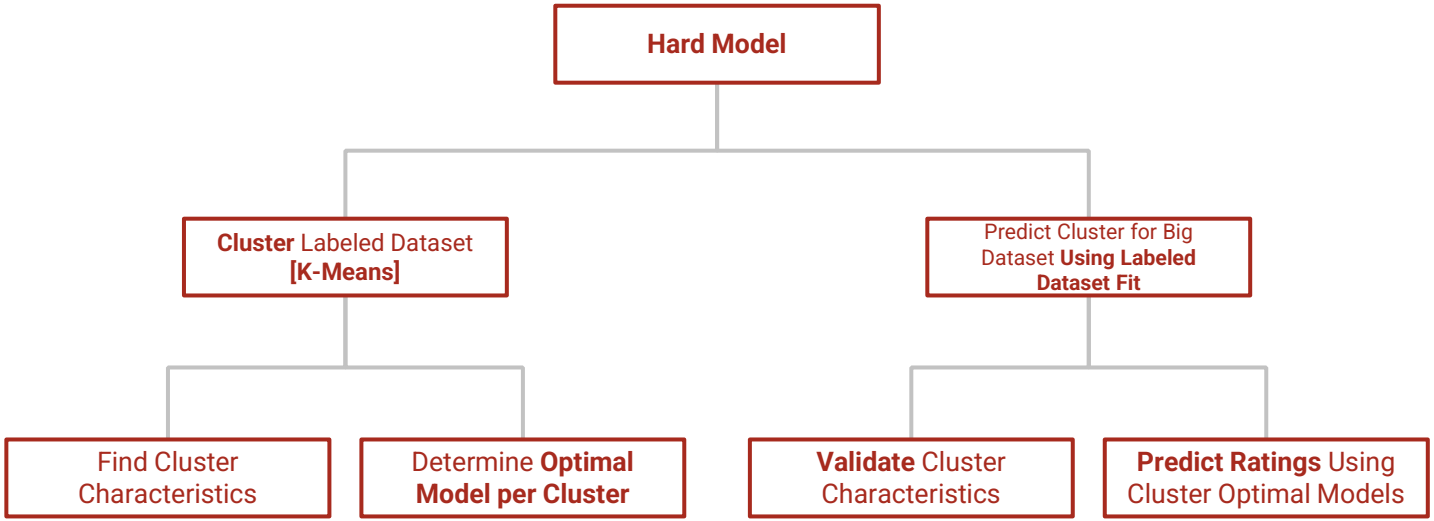
4. Clustering & Segmentation

Hard Model

Clusterings for Segmentation

Scotiabank® X Rotman

Approach



**Ratings not used as variable for clustering algorithm*

Clusterings for Segmentation

Scotiabank® X Rotman

Hard Model

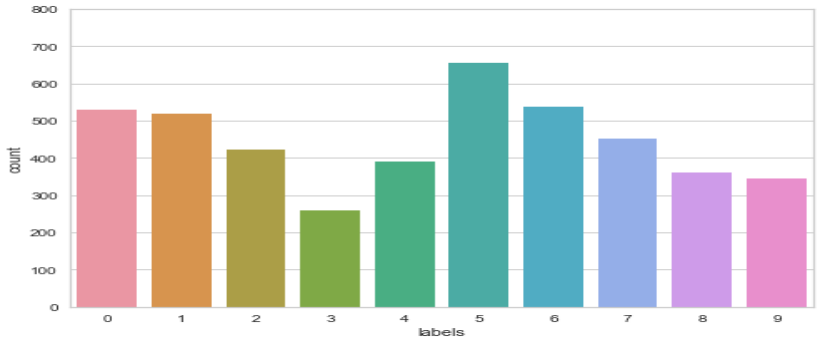
Model Overview: Clustering Model

- **Unsupervised learning** algorithm where there is no target variable.
- K=10 from elbow method criterion.
- Both **BIRCH and K-means** were used to create labels and compared to determine which is the best clustering model.
- K-means was also conducted **without the ratings** to compare future clusters.

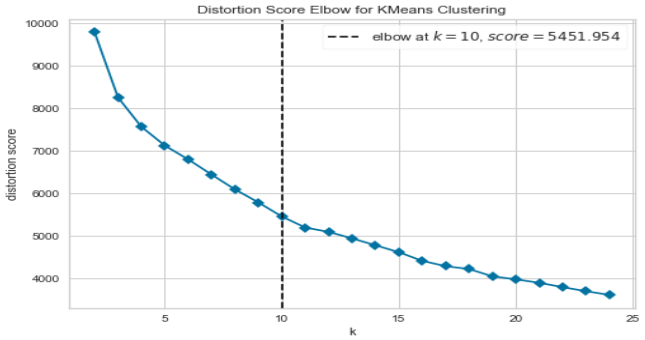
Model Overview: Segmentation Model

- The **optimal model** for each cluster is created from K-means.
- **Accuracy and Recall** rate was utilized as the evaluation metric.
- Models used include Logistic Regression, KNN, Decision Tree Classifier, Random Forest, SVC, and Gradient Boosting Classifier and neural networks
- **10-Fold Cross-Validation** was utilized with 60/40 split

Cluster Label Distribution



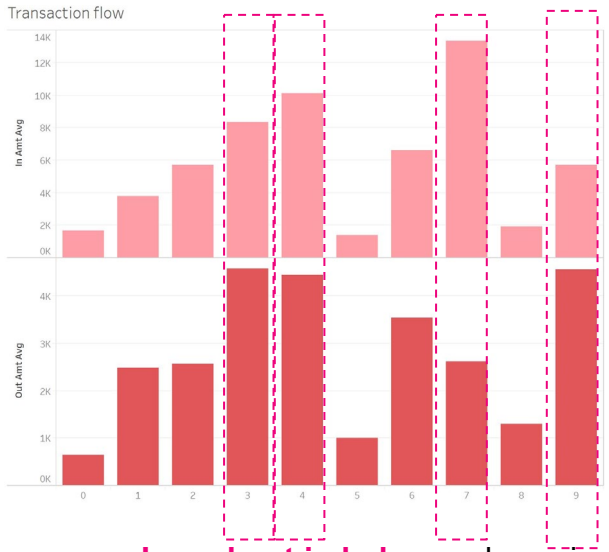
Selecting Optimal K



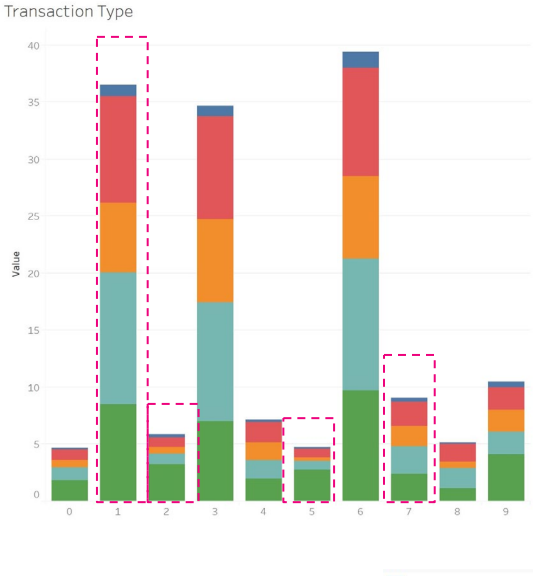
Clusterings for Segmentation

Scotiabank® X Rotman

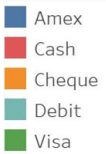
Cluster Characteristics

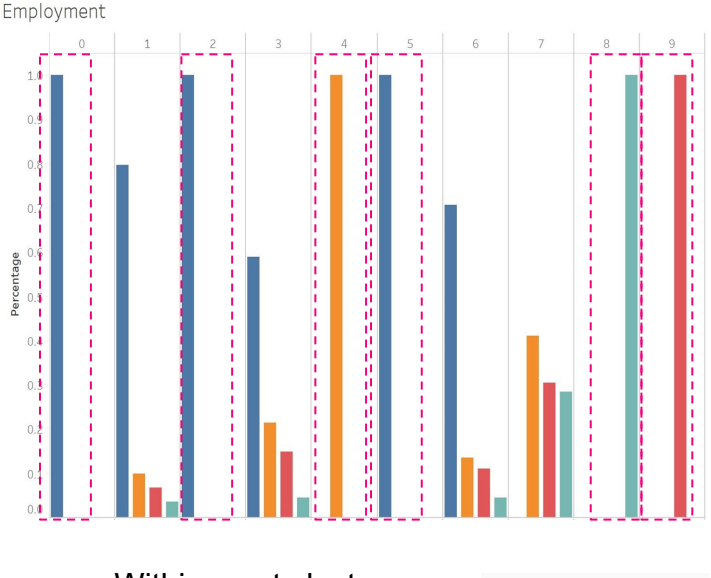


- **In and out imbalance** shown in some clusters
- Some clusters show **consistent** transaction behaviours while others have great volatility

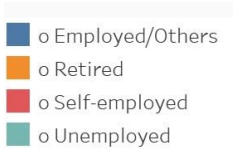


- **Visa usage** is particularly high in several clusters





- Within most clusters, customers either have only **one type** of employment or a **dominant type**.

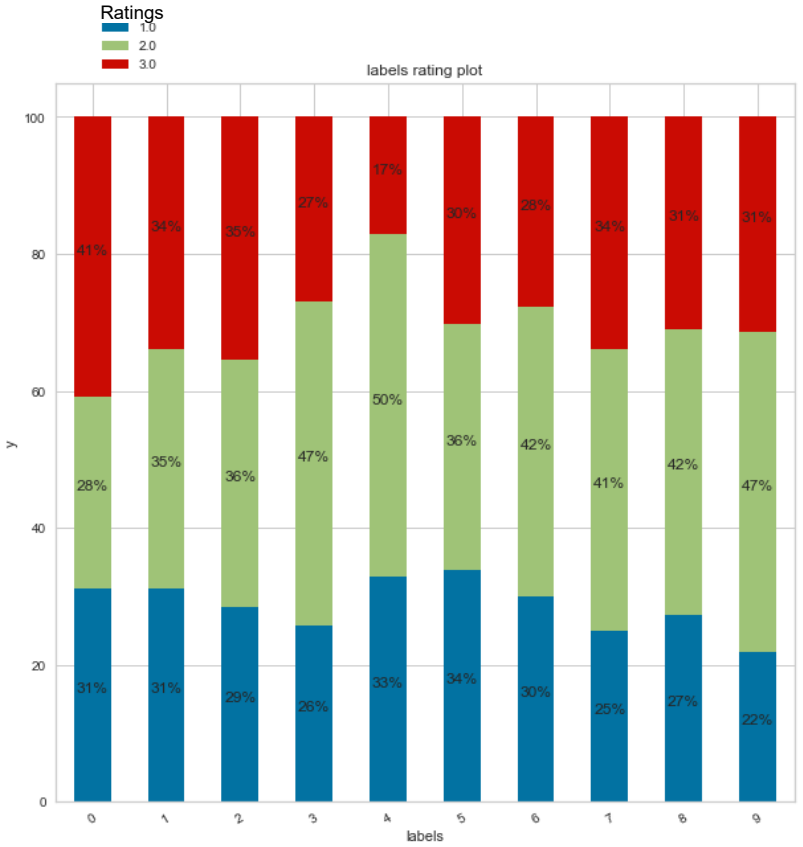


Clusterings for Segmentation

Scotiabank® X Rotman

Characteristics Summary

- Potential **high risk money laundering related** characteristics:
  - High percentage usage of visa (cluster 0,9)
  - High percentage of credit cards (cluster 7,9)
  - Self-employed (cluster 9)
  - Joint accounts (cluster 7)
  - High positive average net incoming amount (incoming amount minus outgoing amount) (cluster 7)
  - High but inconsistent outcoming amount (cluster 9)
  - High but inconsistent incoming amount (cluster 7)
- Potential **low risk money laundering related** characteristics:
  - Active accounts (cluster 4)
  - Retiree (cluster 4)
  - Owner relationship type (cluster 6)
  - High percentage of term deposits (cluster 6)

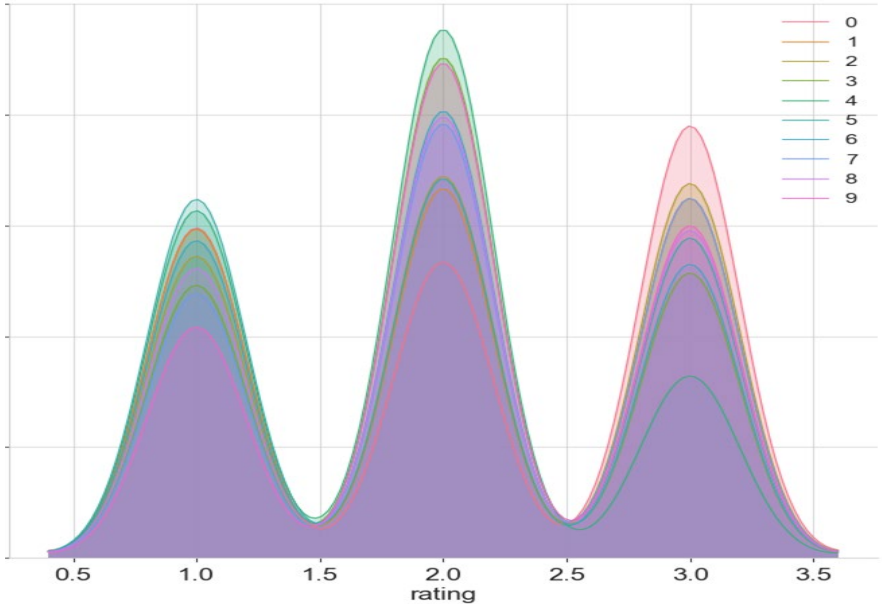


Clusterings for Segmentation

Scotiabank® X Rotman

Cluster Profiles

Rating Distribution



Cluster Group Labels

High Potential Risk Group

0 7 9

Medium Potential Risk Group

1 2 3 8

Lower Potential Risk Group

4 5 6

Clusterings for Segmentation

Scotiabank® X Rotman

Optimal Models

Optimal Models Accuracy Rate Overview

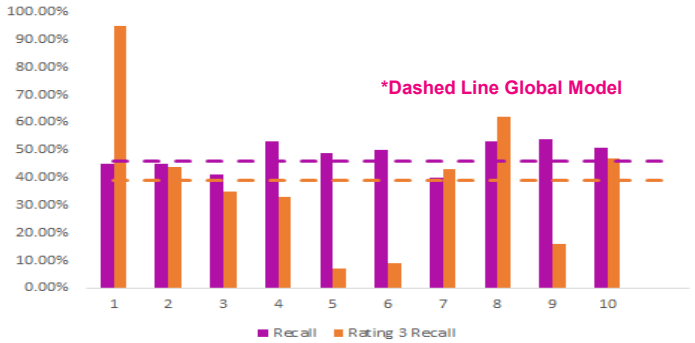
Clusters	Logistic Regression	Decision Tree	KNN	SVC	Gradient Boosting Classifier	Random Forest	ANN	N
Cluster 0	0.373	0.396	0.354	0.377	0.420	0.448	0.260	212
Cluster 1	0.409	0.337	0.346	0.418	0.447	0.418	0.327	208
Cluster 2	0.385	0.385	0.367	0.402	0.385	0.408	0.367	169
Cluster 3	0.510	0.385	0.385	0.471	0.500	0.529	0.5	104
Cluster 4	0.494	0.481	0.449	0.442	0.455	0.481	0.461	156
Cluster 5	0.485	0.466	0.450	0.504	0.454	0.466	0.328	262
Cluster 6	0.386	0.400	0.386	0.386	0.377	0.400	0.386	215
Cluster 7	0.511	0.412	0.473	0.489	0.451	0.533	0.407	182
Cluster 8	0.510	0.524	0.538	0.531	0.531	0.545	0.4	145
Cluster 9	0.449	0.413	0.406	0.507	0.406	0.478	0.442	138

Model Performance Metrics

Test Accuracy and **Recall** on Rating Target Variable

Model Interpretation

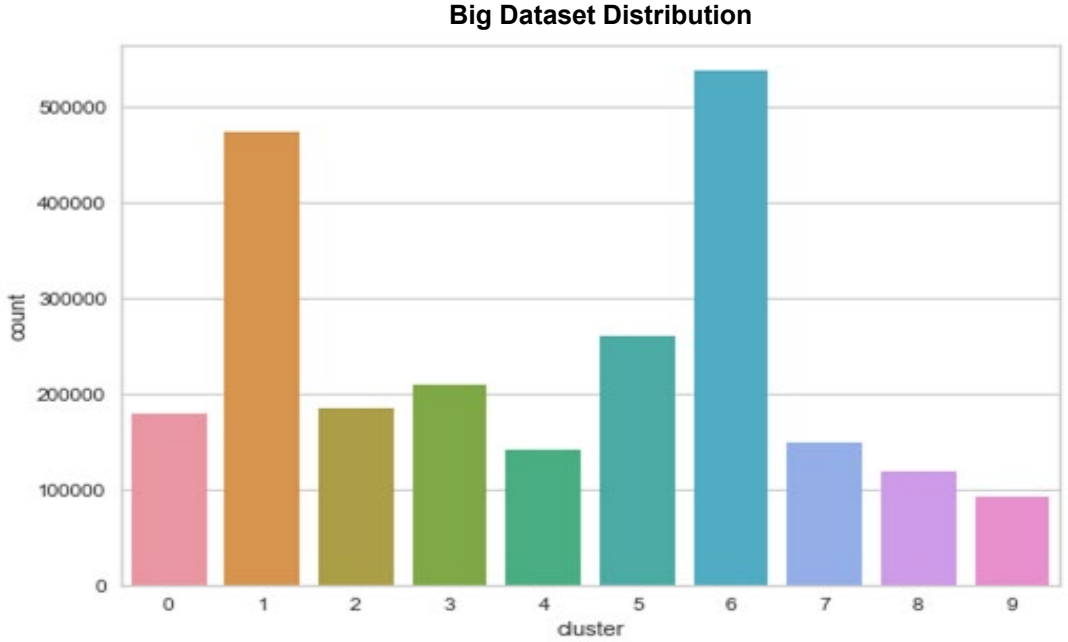
- **Global model** test accuracy was: 45.1% with the best model being Gradient Boosting Classifier.
- The **Weighted Test Set Accuracy** was 47.77 %.
- With cluster labels being **validated** we can **assume similar performance** with the giant dataset as we would with the smaller datasets.



Clusterings for Segmentation

Rotman

Predicting Labels on Big Dataset



Big Dataset

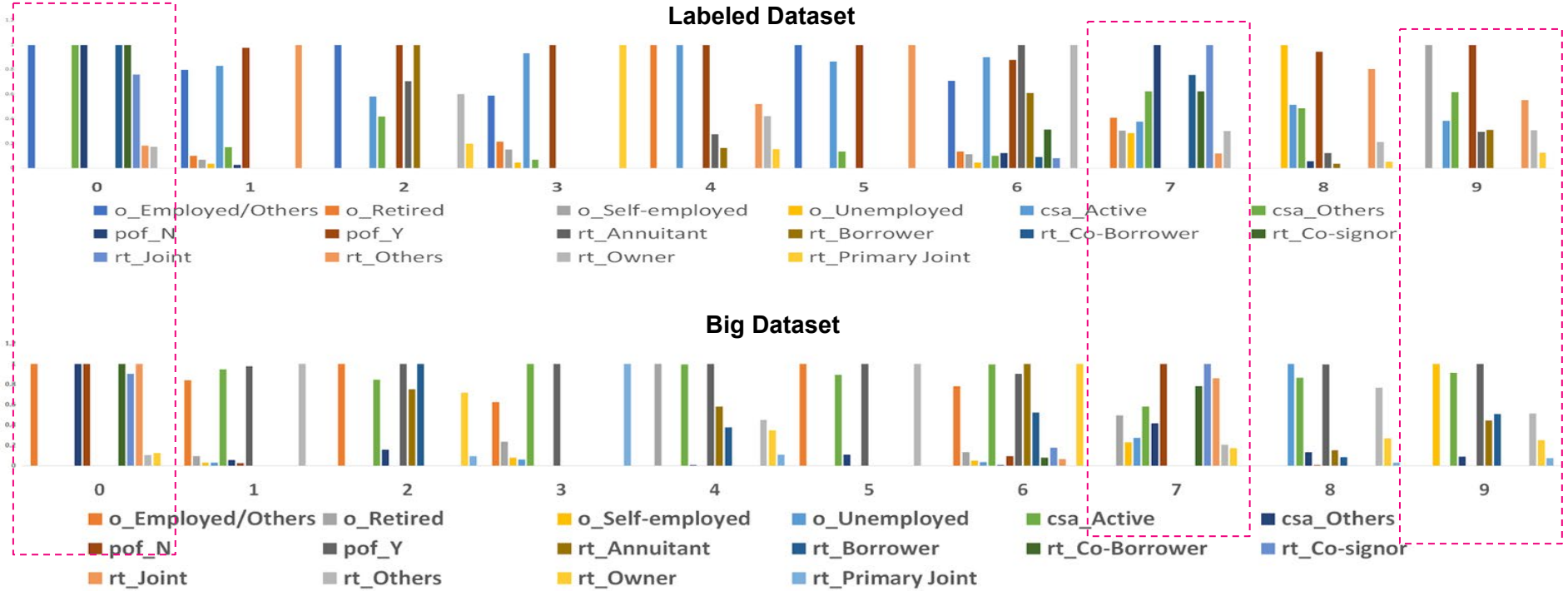
- **Using K-Means algorithm fit on the labeled dataset**, we predict the cluster labels for the big dataset.
- **Assumption:** If the underlying characteristics of clusters between the two datasets are the same, then optimal models are valid to use.

Clusterings for Segmentation

Scotiabank® X Rotman

Validation between Labeled Dataset and Big Dataset

- Labeled Dataset and Big Dataset Clusters demonstrate **similar comparative patterns.**



6. Recommendation & Insights

Summary

Recommendation & Insights

Scotiabank® X Rotman

Overall Findings & Recommendation

Model Findings

Easy and Medium Model provides…

- High accuracy on prediction

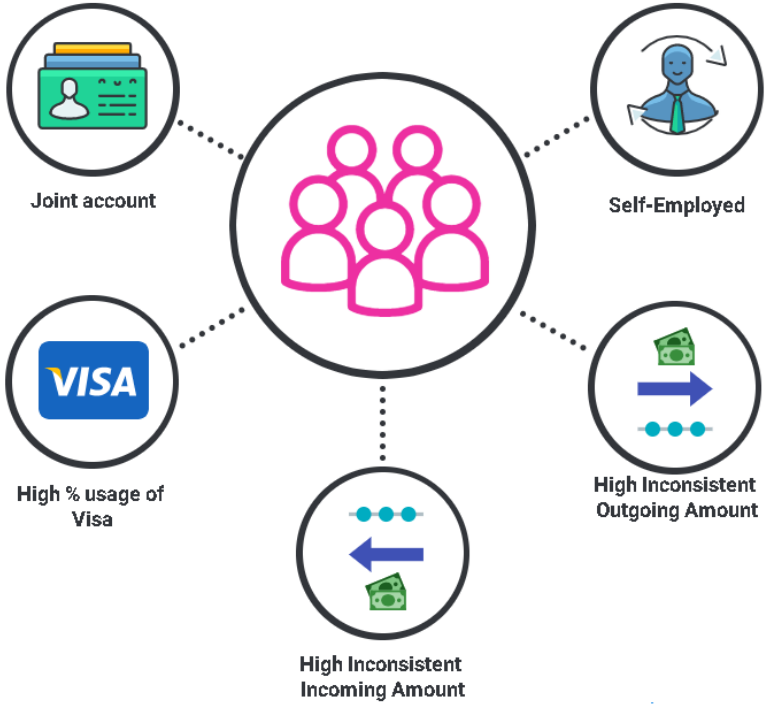
But…

- **Unable to validate** on big data
- Low interpretability

Segmentation Model Improves…

- Accuracy: 45.1% → **47.8%**
- **Interpretability** for insights

Feature Findings



Recommendation & Insights

Scotiabank® X Rotman

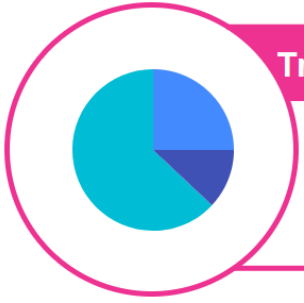
Binary Classification

Make target variable binary where prediction is if it is high risk or not (level 3 or not)

- Accuracy improves to ~70%+.

More Labeled Datasets

Create another labeled dataset to test the segmented model approach.



Track Customer Clusters

Track customers that are in the high potential risk group clusters, monitor changes to develop trends.

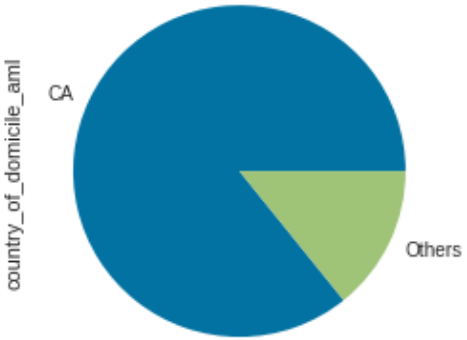
Thank you

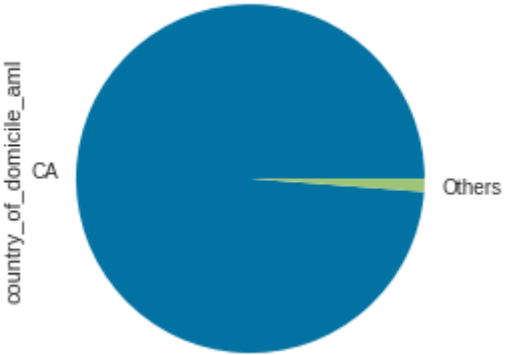
Questions?

Country Code Distribution

Rotman

Figure A1: (Left) Country Distribution for Training (Right) Country Distribution for Testing





Transaction Data

Rotman

Table 1: Aggregate measures for transaction data

Original Columns	Incoming Amount Incoming Count	Outgoing Amount Outgoing Count	Others
Aggregated Columns	Count, Average, Max, Min, Standard Deviation	Count, Average, Max, Min, Standard Deviation	Number of Month without incoming or outgoing transaction, Number of transaction type used Usage Count for each type of transaction

Easy Model

Rotman

Pipeline - Results & Hyperparameters

Pipeline Models	Training Accuracy	Test Accuracy	Precision	Recall	F-1 score
Logistic Regression	**0.429**	**0.451**	**0.450**	**0.450**	**0.430**
Decision Tree	**0.431**	**0.444**	**0.320**	**0.440**	**0.370**
KNN	**0.413**	**0.429**	**0.440**	**0.430**	**0.430**
SVC	**0.434**	**0.450**	**0.440**	**0.450**	**0.440**
Random Forest	**0.448**	**0.456**	**0.440**	**0.460**	**0.440**
Gradient Boosting	**0.451**	**0.460**	**0.450**	**0.460**	**0.460**

```
Estimator: Gradient Boosting Classifier
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'deviance', 'clf__n_estimators': 100}
Best training accuracy: 0.451
Test set accuracy score for best params: 0.460
              precision    recall  f1-score   support

         1.0       0.36      0.34      0.35       371
         2.0       0.53      0.60      0.56       548
         3.0       0.44      0.38      0.41       422

    accuracy                           0.46      1341
   macro avg       0.44      0.44      0.44      1341
weighted avg       0.45      0.46      0.46      1341
```

	precision	recall	f1-score	support	pred	AUC
1.0	0.371831	0.355795	0.363636	371.0	355.0	0.609941
2.0	0.549669	0.605839	0.576389	548.0	604.0	0.674856
3.0	0.429319	0.388626	0.407960	422.0	382.0	0.639221
avg / total	0.462596	0.468307	0.464526	1341.0	1341.0	0.651864

Lasso Regression

Medium Model Predictors

Predictor	Coefficients
j_Others	-1.211907e+00
csa_Others	-1.155181e+00
rt_Owner	-4.900006e-01
PCD_CDA	-3.799143e-01
o_Unemployed	-3.701852e-01
rt_Co-signor	3.541128e-01
c_0	-3.215311e-01
trsactn_type_count	-3.091671e-01
o_Self-employed	-2.925306e-01
rt_Annuitant	-1.927481e-01
rt_Primary Joint	-1.891245e-01
PCD_CRC	1.407186e-01
in_cnt_min	1.066918e-01
PCD_TED	9.156492e-02
o_Retired	7.737355e-02
rt_Borrower	-7.075788e-02
PCD_MOR	-7.024898e-02
pof_N	-6.558559e-02
out_cnt_min	-5.871862e-02
visa	5.303453e-02
PCD_SAV	4.952730e-02
debit	4.075089e-02
cheque	3.517524e-02
PCD_LLC	3.247072e-02
cash	3.208445e-02
in_cnt_avg	-2.107545e-02
out_cnt_null	2.015289e-02
out_cnt_avg	1.706504e-02
in_cnt_std	1.630816e-02
rt_Joint	1.501535e-02
out_cnt_std	-1.498699e-02
in_cnt_sum	-4.615618e-03
in_cnt_null	-4.344357e-03
amex	-4.142899e-03
rt_Co-Borrower	3.408981e-03
out_cnt_max	-1.307428e-03
out_cnt_sum	-2.487149e-04
out_amt_avg	2.995869e-05
out_amt_min	-2.704371e-05
out_amt_std	-1.835016e-05
in_amt_std	-1.262050e-05
in_amt_avg	9.480494e-06
in_amt_max	3.706474e-06
in_amt_min	-2.574449e-06
out_amt_max	1.726694e-06
in_amt_sum	-8.511916e-07
out_amt_sum	-4.903434e-07

Figure D2: Feature Selection with Lasso Regression

Medium Model

Rotman

Pipeline - Results & Hyperparameters

Pipeline Models	Training Accuracy	Test Accuracy	Precision	Recall	F-1 score
Logistic Regression	**0.449**	**0.452**	**0.440**	**0.450**	**0.440**
Decision Tree	**0.456**	**0.438**	**0.440**	**0.440**	**0.440**
KNN	**0.431**	**0.434**	**0.430**	**0.430**	**0.420**
SVM	**0.461**	**0.457**	**0.450**	**0.460**	**0.450**
Random Forest	**0.480**	**0.444**	**0.430**	**0.440**	**0.420**
Gradient Boosting	**0.456**	**0.456**	**0.450**	**0.460**	**0.450**

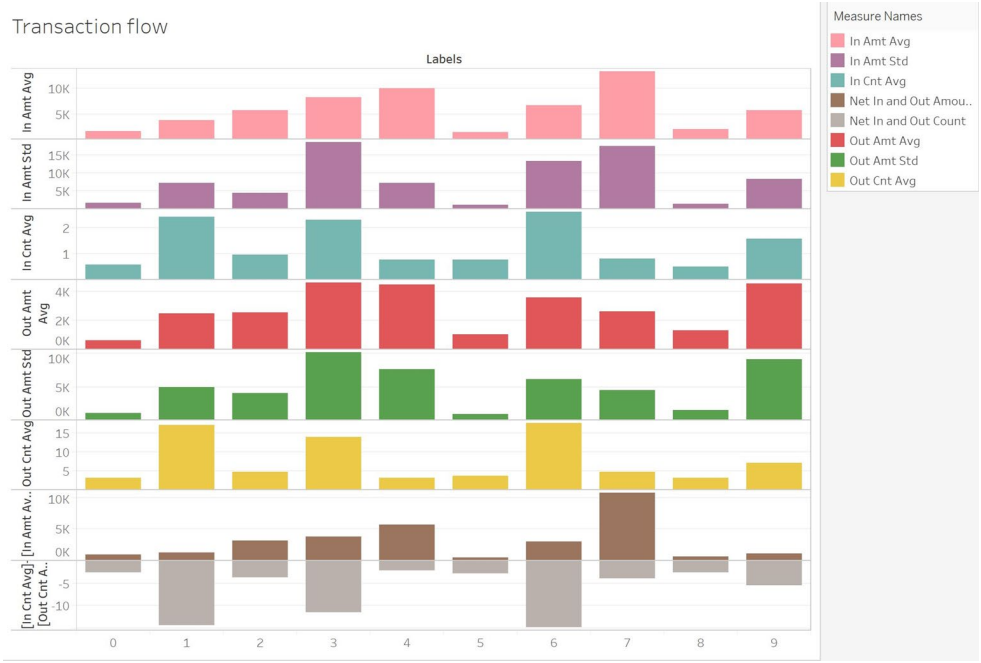
```
Estimator: Gradient Boosting Classifier
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'deviance', 'clf__n_estimators': 100}
Best training accuracy: 0.461
Test set accuracy score for best params: 0.456
              precision    recall  f1-score   support

         1.0       0.40      0.32      0.36       545
         2.0       0.50      0.61      0.55       684
         3.0       0.44      0.39      0.41       559

    accuracy                           0.46      1788
   macro avg       0.44      0.44      0.44      1788
weighted avg       0.45      0.46      0.45      1788
```

	precision	recall	f1-score	support	pred	AUC
1.0	0.392936	0.326606	0.356713	545.0	453.0	0.610482
2.0	0.504294	0.600877	0.548366	684.0	815.0	0.671934
3.0	0.432692	0.402504	0.417053	559.0	520.0	0.640295
avg / total	0.447966	0.455257	0.448895	1788.0	1788.0	0.646521

Cluster Characteristics

Rotman

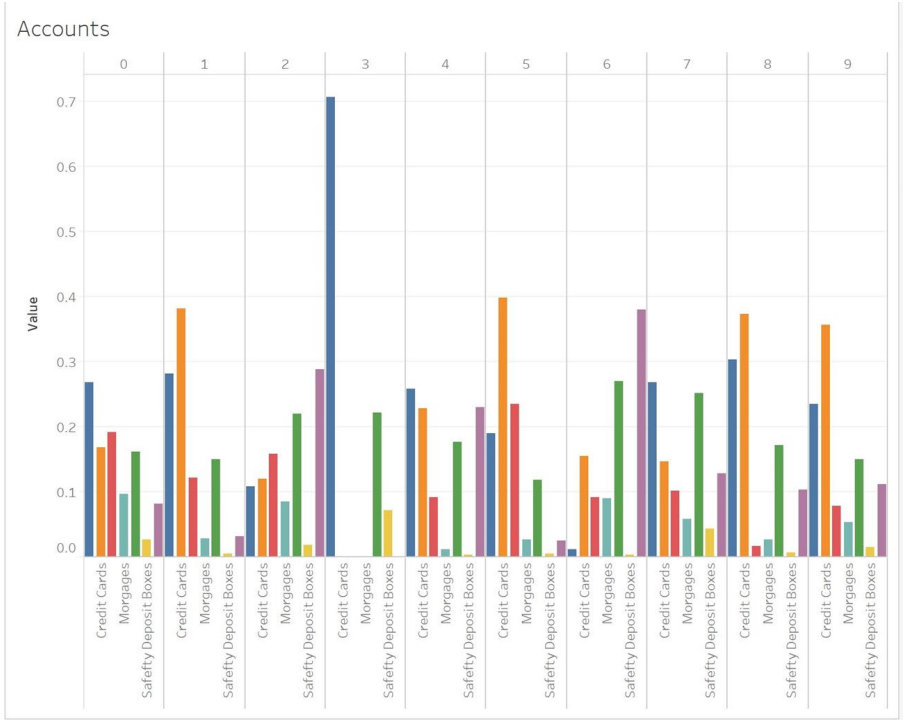
Figure F1: Transaction Flow Diagram



Clusterings for Segmentation

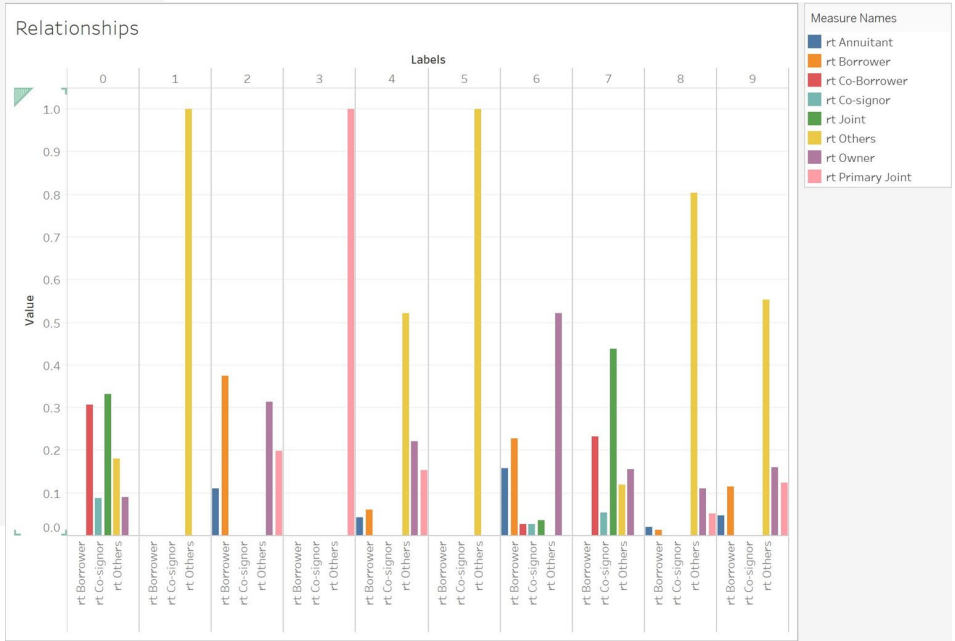
Scotiabank® X Rotman

Cluster Characteristics



- Most accounts type is either chequing, credit card or term deposit.

- Different relationship types among clusters, dominant ones are obvious

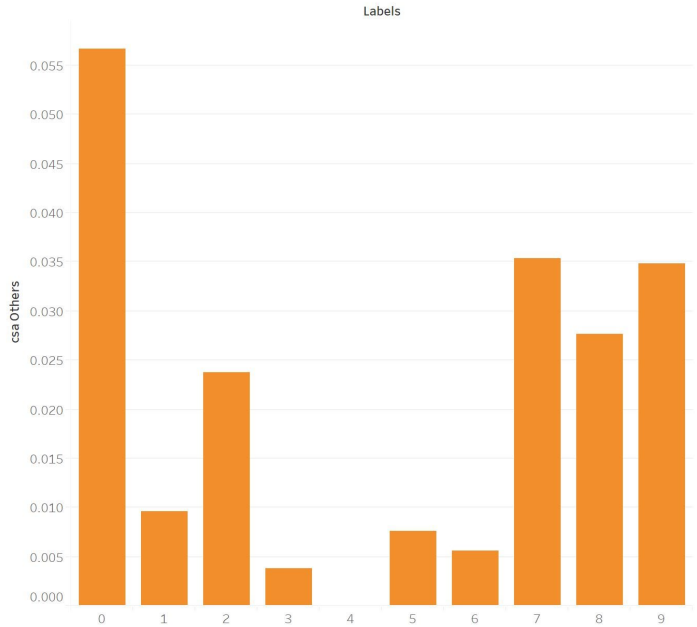


Clusterings for Segmentation

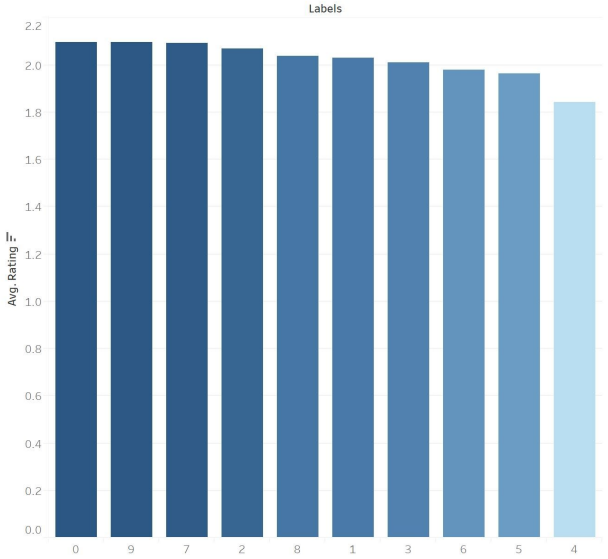
Scotiabank® X Rotman

Summary

Non-active Accounts



Average Rating



Cluster Characteristics Summary

Rotman

Cluster	0	1	2	3	4	5	6	7	8	9
Transaction flow	Low average and std of in and out amount	High in and out counts	Positive in amount	High in and out counts	Low in and out counts	Low average and std of in and out amount	High in and out counts	High but inconsistent in amount	Low average and std of in and out amount	High and inconsistent out amount
	Low in and out counts	Low in and out amount		High average and std in and out amount	High in and out amount	Low in and out counts	Positive in amount	Low in and out counts	Low in and out counts	Neutral in and out amount
	Neutral in and out amount	Neutral in and out amount		Positive in amount	Positive in amount	Neutral in and out amount	More out counts	High Positive in amount	Neutral in and out amount	
		More out counts		More out counts						
Employment	Only have employed		Only have employed		Only have Retired	Only have employed			Only have unemployed	Only have self-employed
Transaction Type	High Percentage of Visa		High Percentage of Visa		Averaged	High Percentage of Visa			Relatively high in Cash and Debit	High Percentage of Visa
			Low in cheques			Low in cheques				
Accounts		High credit cards		High chequing		High credit cards	High term deposits		High credit cards	High credit cards
Relationships		Only Others	Borrowers/Owners	Only Primary Joint		Only Others	Owner	Joints	Others	Others
Customer Status	Mostly Inactive				All active					

Hard Model Model Tuning

Rotman

Cluster 1

```
Estimator: Random Forest
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 10, 'clf__min_impurity_decrease': 0.01, 'clf__min_samples_split': 40}
Best training accuracy: 0.365
Test set accuracy score for best params: 0.448
              precision    recall  f1-score   support

         1.0       0.29      0.03      0.06        62
         2.0       0.50      0.05      0.10        55
         3.0       0.45      0.95      0.61        95

    accuracy                           0.45       212
   macro avg       0.41      0.34      0.26       212
weighted avg       0.42      0.45      0.32       212


Classifier with best test set accuracy: Random Forest
```

Cluster 2

```
Estimator: Gradient Boosting Classifier
Best params: {'clf__learning_rate': 1.0, 'clf__loss': 'deviance', 'clf__n_estimators': 300}
Best training accuracy: 0.410
Test set accuracy score for best params: 0.447
              precision    recall  f1-score   support

         1.0       0.39      0.46      0.42        61
         2.0       0.46      0.44      0.45        68
         3.0       0.49      0.44      0.46        79

    accuracy                           0.45       208
   macro avg       0.45      0.45      0.45       208
weighted avg       0.45      0.45      0.45       208
```

Hard Model Model Tuning

Rotman

Cluster 3

```
Estimator: Random Forest
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 20, 'clf__min_impurity_decrease': 0.01, 'clf__min_samples_split': 20}
Best training accuracy: 0.413
Test set accuracy score for best params: 0.408
              precision    recall  f1-score   support

         1.0       0.28      0.17      0.21        41
         2.0       0.47      0.63      0.54        62
         3.0       0.38      0.35      0.36        66

    accuracy                           0.41       169
   macro avg       0.38      0.38      0.37       169
weighted avg       0.39      0.41      0.39       169


Classifier with best test set accuracy: Random Forest
```

Cluster 4

```
Estimator: Random Forest
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 10, 'clf__min_impurity_decrease': 0.01, 'clf__min_samples_split': 20}
Best training accuracy: 0.488
Test set accuracy score for best params: 0.529
              precision    recall  f1-score   support

         1.0       0.29      0.16      0.21        25
         2.0       0.58      0.81      0.68        52
         3.0       0.50      0.33      0.40        27

    accuracy                           0.53       104
   macro avg       0.46      0.43      0.43       104
weighted avg       0.49      0.53      0.49       104


Classifier with best test set accuracy: Random Forest
```

Hard Model Model Tuning

Rotman

Cluster 5

```
Estimator: Logistic Regression
Best params: {'clf__C': 1.0, 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}
Best training accuracy: 0.538
Test set accuracy score for best params: 0.494
              precision    recall  f1-score   support

         1.0       0.45      0.24      0.31        54
         2.0       0.51      0.86      0.64        72
         3.0       0.40      0.07      0.11        30

    accuracy                           0.49       156
   macro avg       0.45      0.39      0.36       156
weighted avg       0.47      0.49      0.43       156
```

Cluster 6

```
Estimator: SVC
Best params: {'clf__C': 1, 'clf__kernel': 'linear'}
Best training accuracy: 0.473
Test set accuracy score for best params: 0.504
              precision    recall  f1-score   support

         1.0       0.45      0.87      0.59        90
         2.0       0.61      0.53      0.57        86
         3.0       0.57      0.09      0.16        86

    accuracy                           0.50       262
   macro avg       0.55      0.50      0.44       262
weighted avg       0.54      0.50      0.44       262
```

Hard Model Model Tuning

Rotman

Cluster 7

```
Estimator: Decision Tree
C:\Users\buttb\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1221: UndefinedMetricWarning: Precision and F-sco
re are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this beha
vior.
  _warn_prf(average, modifier, msg_start, len(result))
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 10, 'clf__min_impurity_decrease': 0.01, 'clf__min_samples_split': 20}
Best training accuracy: 0.410
Test set accuracy score for best params: 0.400
              precision    recall  f1-score   support

         1.0       0.36      0.14      0.20        64
         2.0       0.44      0.58      0.50        83
         3.0       0.36      0.43      0.39        68

    accuracy                           0.40       215
   macro avg       0.39      0.38      0.36       215
weighted avg       0.39      0.40      0.38       215
```

Cluster 8

```
Estimator: Random Forest
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 20, 'clf__min_impurity_decrease': 0.001, 'clf__min_samples_split': 6
0}
Best training accuracy: 0.509
Test set accuracy score for best params: 0.533
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00        40
         2.0       0.50      0.74      0.60        74
         3.0       0.58      0.62      0.60        68

    accuracy                           0.53       182
   macro avg       0.36      0.45      0.40       182
weighted avg       0.42      0.53      0.47       182
```

Hard Model Model Tuning

Rotman

Cluster 9

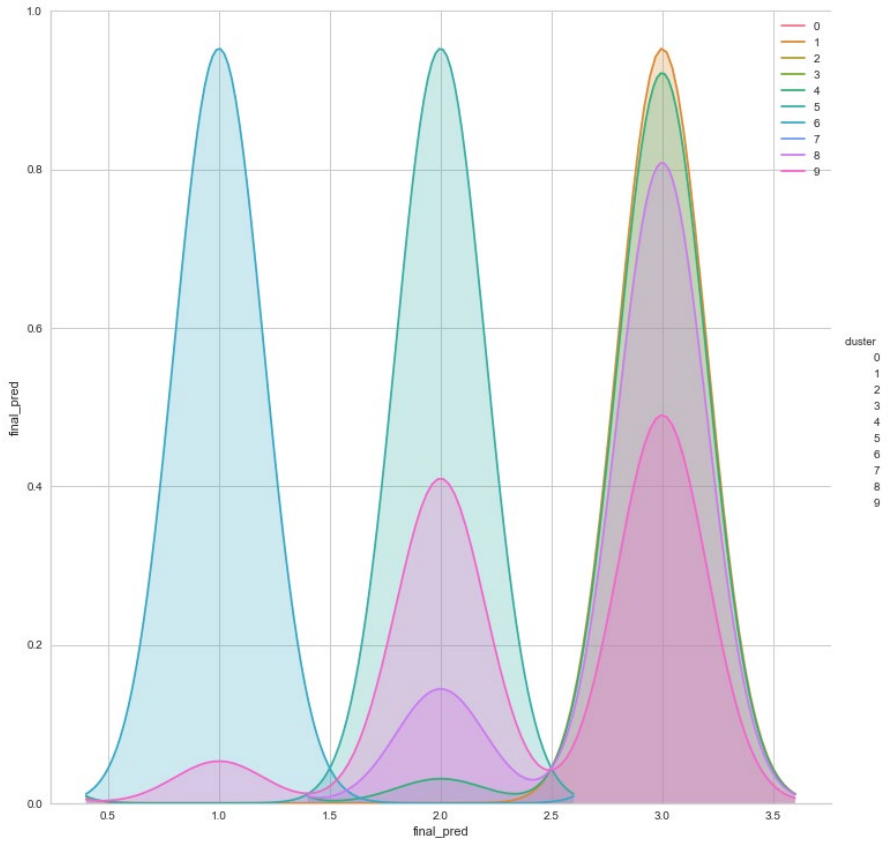
```
Estimator: Random Forest
Best params: {'clf__criterion': 'gini', 'clf__max_depth': 10, 'clf__min_impurity_decrease': 0.005, 'clf__min_samples_split': 6
0}
Best training accuracy: 0.539
Test set accuracy score for best params: 0.545
              precision    recall  f1-score   support

         1.0       0.46      0.55      0.50        38
         2.0       0.56      0.86      0.68        58
         3.0       0.89      0.16      0.28        49

    accuracy                           0.54       145
   macro avg       0.63      0.53      0.48       145
weighted avg       0.64      0.54      0.49       145
```

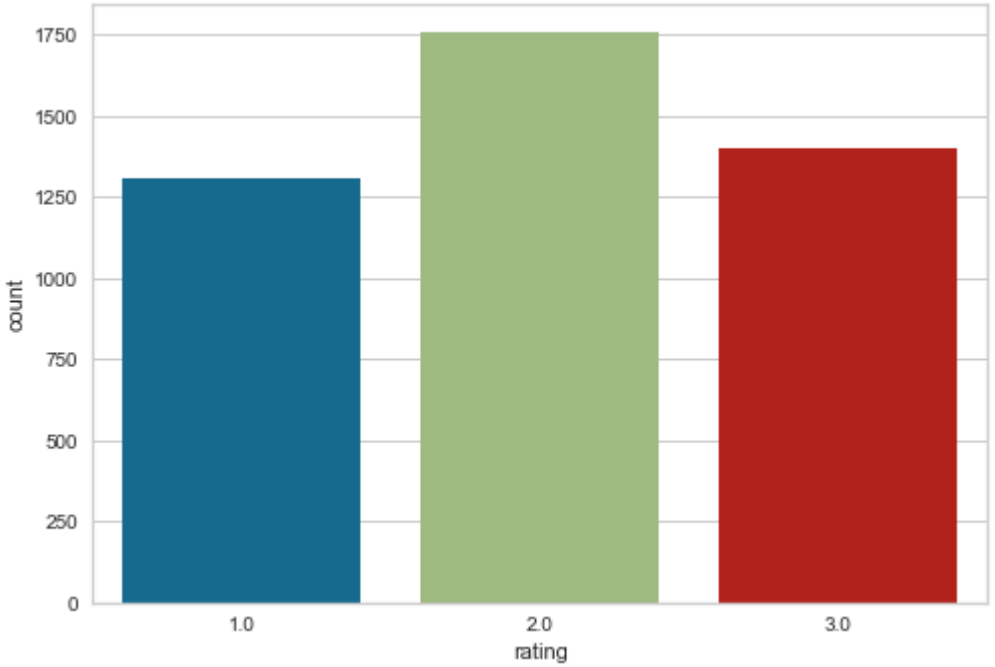
Cluster 10

```
Estimator: SVC
Best params: {'clf__C': 10, 'clf__kernel': 'rbf'}
Best training accuracy: 0.513
Test set accuracy score for best params: 0.507
              precision    recall  f1-score   support

         1.0       0.40      0.25      0.31        32
         2.0       0.55      0.67      0.61        61
         3.0       0.48      0.47      0.47        45

    accuracy                           0.51       138
   macro avg       0.48      0.46      0.46       138
weighted avg       0.49      0.51      0.49       138
```

Rating Predictors Big Dataset

Rotman




1750
1500
1250
1000
750
500
250
0
count
1.0
2.0
3.0
rating