Bank of Crown

Data Modelling

YOUNG PROFESSIONAL PROGRAM

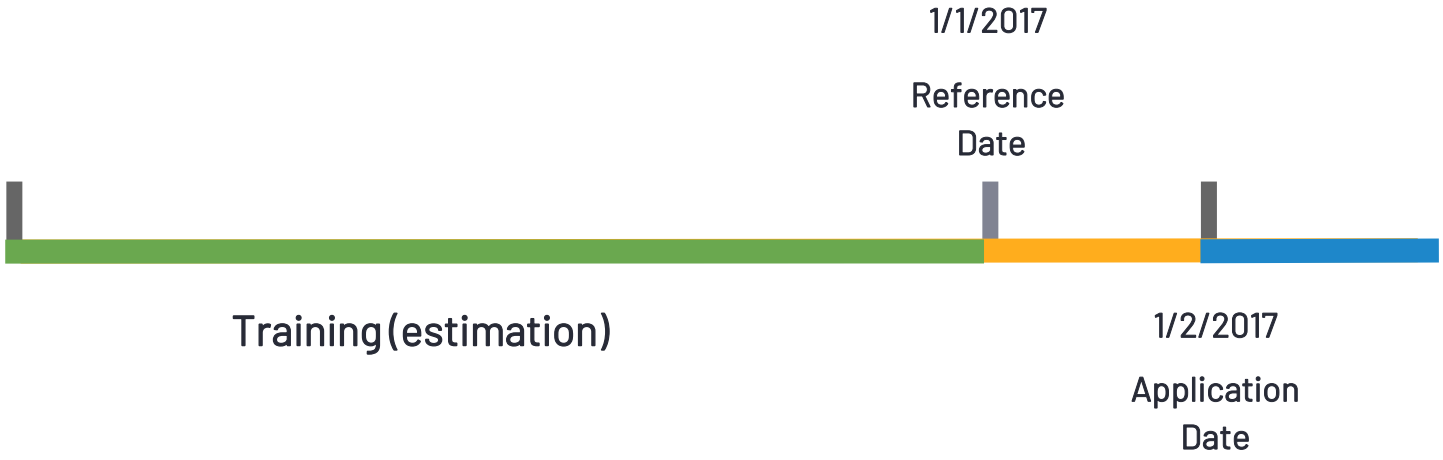
The Bank of Crown

The Bank of crown has 3 key areas to focus on:

▪ Reducing Risk from Bank Loans	Classification	SAP Predictive Analytics
▪ Enhancing Communications Strategy	Clustering	
▪ Improving Customer Satisfaction	Regression	Predictive Factory

The Bank of Crown

Training the model on the previous 6 months and applying it with a latency of one month to allow appropriate reaction



Data Understanding

Description, Exploration & Quality

YOUNG PROFESSIONAL PROGRAM

Data Collection

Location

HANA DataBase

Issues?

None

Acquisition

Data Manager
(Predictive Analytics)

Data Description

Accounts

22,500 records
4 Fields

Clients

26,845 records
4 Fields

Disposition

26,845 records
4 Fields

Demographic

77 records
16 Fields

Order

32,355 records
6 Fields

Transactions

5,281,600 records
10 Fields

Loan

3,410 records
7 Fields

Credit Cards

4,460 records
4 Fields

Data Exploration

Interesting Findings

1. 83% of accounts do not have **credit card** issued
2. Only 16% of account are requesting loans and only 3% are Defaulters
3. Loan month **duration** has four unused categories
4. Loan **amount** ranges from 4980 to 590820
5. High frequency of classic **clients**
6. **Bank** name does not have any entry

Data Exploration : Initial Model

Predict unpaid loans before approving them

Loan Status as target Field

'A' - contract finished, no problems

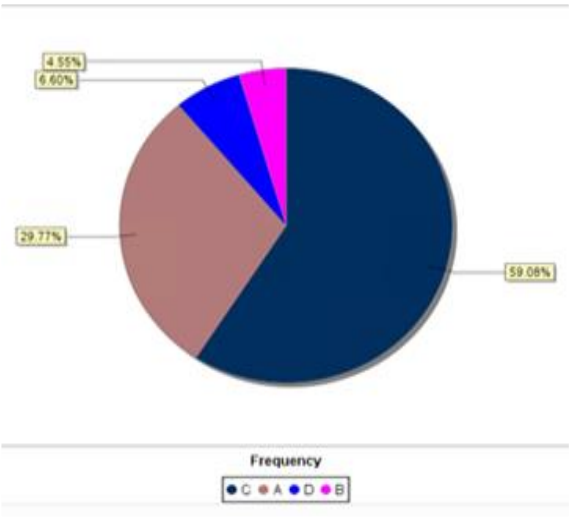
'C' - running contract, OK so far

= 0

'B' - contract finished, loan not paid

'D' - running contract, client in debt

= 1

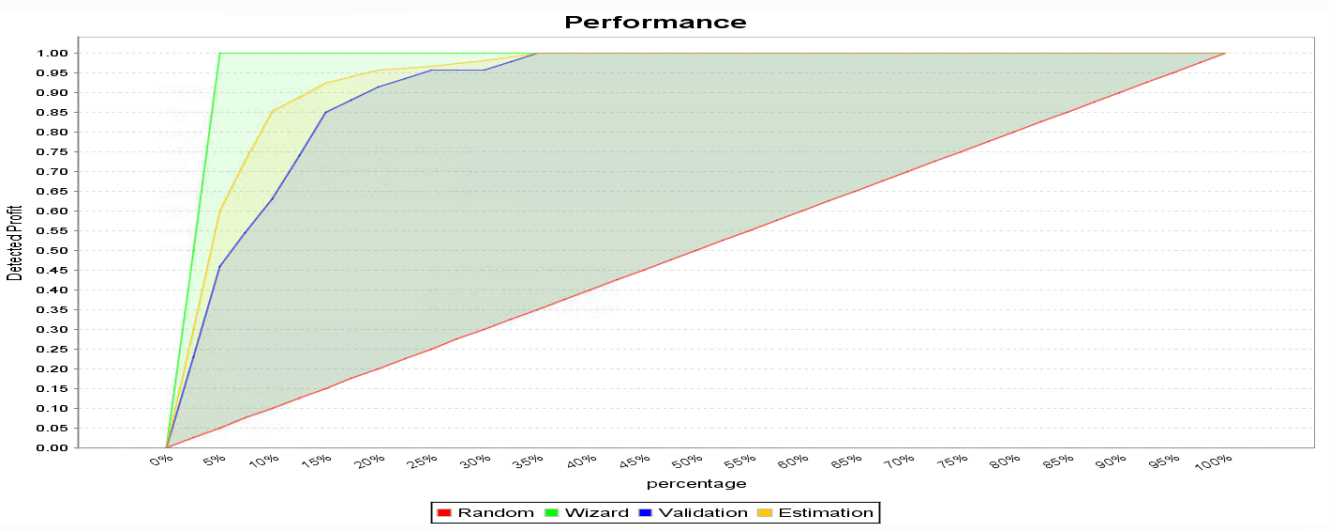


Data Exploration : Initial Model

Predict unpaid loans before approving them

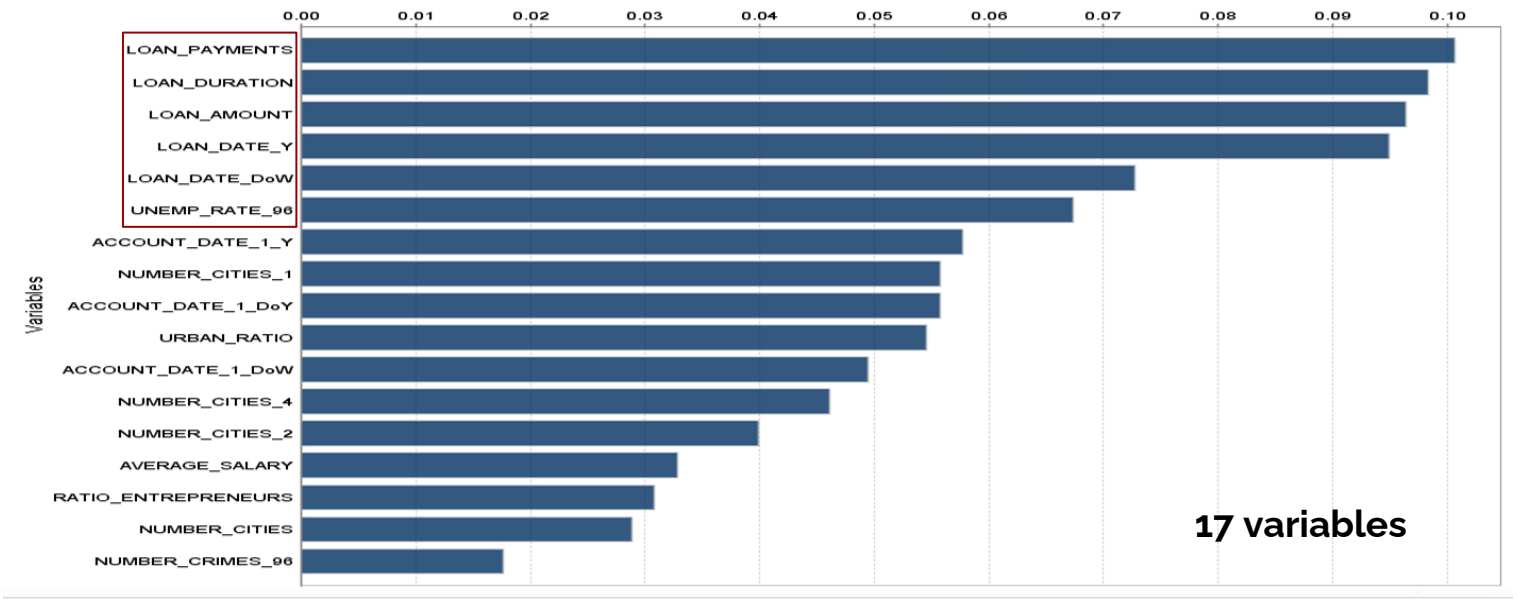
Predictive Power (KI) : 0.8699

Prediction Confidence (KR) :0.9447



Data Exploration : Initial Model

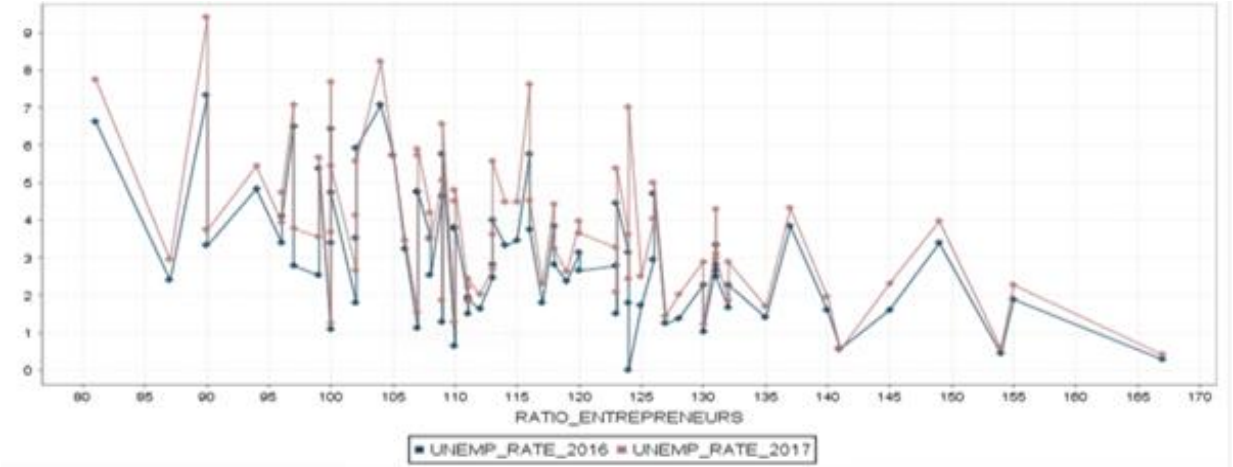
Variable Contribution to Defaulters



Data Exploration : Initial Model

Highly Correlated Variables

Index	First Variable	Second Variable	Coefficient
21	UNEMP_RATE_95	UNEMP_RATE_96	0.984



Data Quality: Missing Values

The variable "**Bank** name" does not has 100% missing value and should be discarded

Table	*Variable*	*Missing Values%*
BOC_ORDERS ***Permanent Orders Data***	Category	20%
BOC_TRANSACTIONS ***(Transactions Data)***	MODE	17.34%
	Type	50.67%
	Bank	**100%**
	Account	70%
BOC_GEODEMO ***(Demographic Data)***	Unemp_rate_2016	6.25%
	Number_crims_2016	6.25%

Data Quality: Outliers

The variable "**District Name**" has 100% outliers which indicates string variables.

Table	*Variable*	*Outliers%*
BOC_TRANSACTIONS ***(Transactions Data)***	AMOUNT	53.53%
BOC_GEODEMO ***(Demographic Data)***	DISTRICT_NAME	**100%**

1. First Scenario

Reducing Risk from Bank Loans

YOUNG PROFESSIONAL PROGRAM

Reduce Risk from Bank Loans (Overview)

Business Goals

This project has the following objectives:

- Increase profits by identifying customers most likely to pay back loans
- Reduce risk by avoiding making loans to potential defaulters

Business Success Criteria

This project will be judged a success if:

- The Bank of the crown reduces the number of bad loans from 11% to 7%

Data Science Goals

- Create a **classification model** that predicts: which current customers might default on a loan
- Latency period:
  - No latency
- History period:
  - 6 months
- Target Period:
  - 1 month
- Population filters: only include those customers who have loans
- Target Variable: Loan Default (did this customer default yes/no)
- Create customer profile of Defaulters vs Non-defaulters
- Model will be operationalized and applied monthly using Predictive Factory

Data Science Success Criteria

- Predictive Power: > 0.5
- Prediction Confidence: > 0.95
- Variable contributions make business sense
- Model performance in evaluation period corresponds to performance in training period
- Model will be operationalized and applied monthly on the 1st day of the month, using Predictive Factory

Variables Selection

The initial number of variables used in the model are 22 but the best tradeoff is 13 variables which are:

- Loan Duration (e.g. 12 months)
- Minimum Loan Payment Amount
- Average Loan Payment Amount
- Client Type (e.g. Classic)
- Unemployment rate for 2017
- Card age (Days)
- Number of municipalities (2)
- Client age
- Number of crimes in 2017 (Client area)
- Number of inhabitants in the region (Client)
- Region name (Client)
- Account age (Days)

Ratio of Defaulters

1.34%

Ratio of solvent

98.66%

Model Performance: KI KR

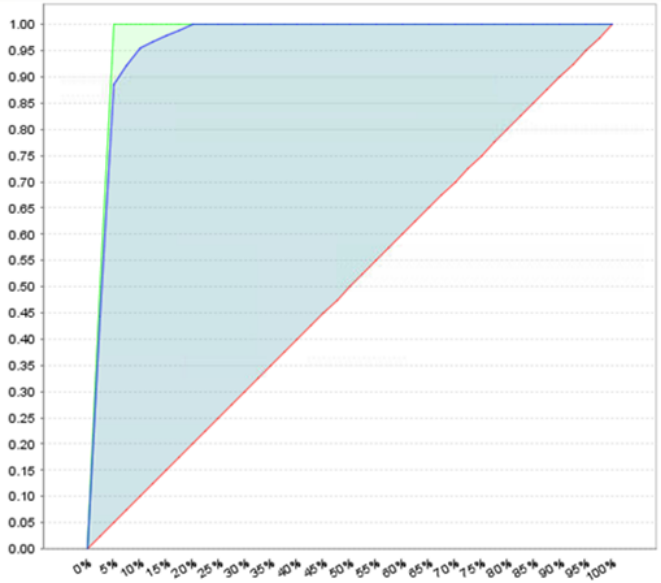
Predictive Power (KI): 97%

Prediction Confidence (KR): 99%

Perfect model
100% correct prediction

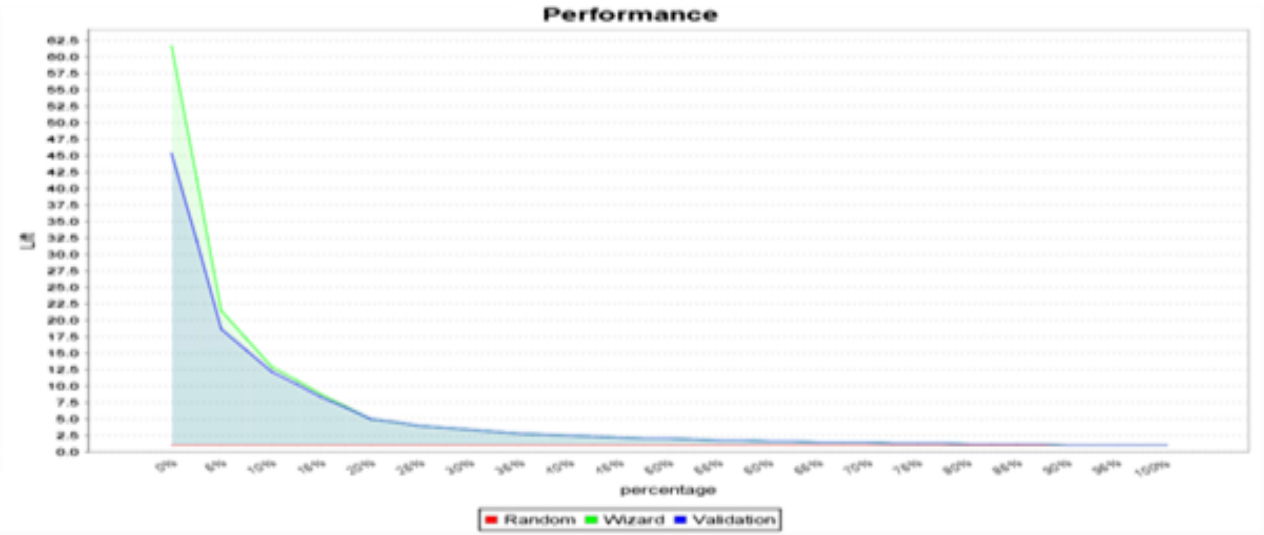
Random model
Worst scenario

Validation
Our model



Model Performance: LIFT-curve

Lift curve of validation in comparison to random prediction



Model Performance: Performance Metric

Metrics

Classification Rate	98.95%
Sensitivity	57.66%
Specificity	99.63%
Precision	71.91%
F1 Score	0.64

By targeting 1% of the clients we detected 58% the entire population of interest, or 17% times better than without a predictive model

Possible Improvement: synthesizing more data

Deviation Analysis

Model performance

Predictive Power (KI): 97%

Prediction Confidence (KR) :99%

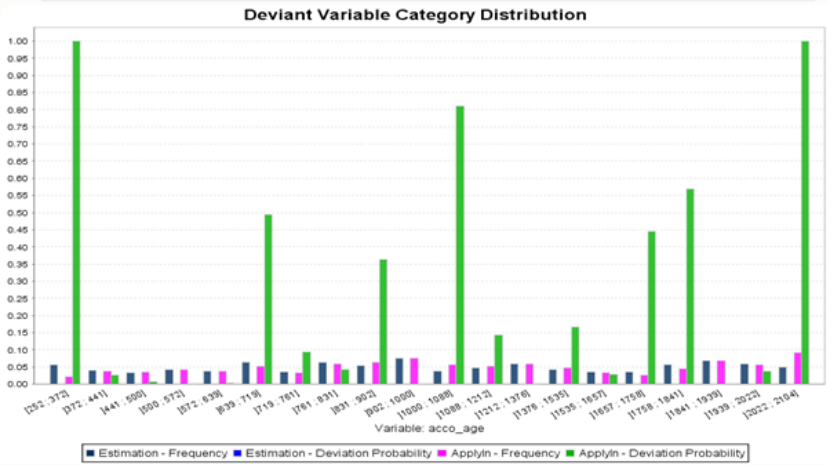
Time sensitive variables:

- Account age (Days)

probability of deviation: 98%

- Card Age (Days)

probability of deviation: 74%



Deviation Analysis

Days → **Categories**

Possible categorization technique:

- Senior
- Mid
- Young

Predictive factory - Deploy (With improvements)

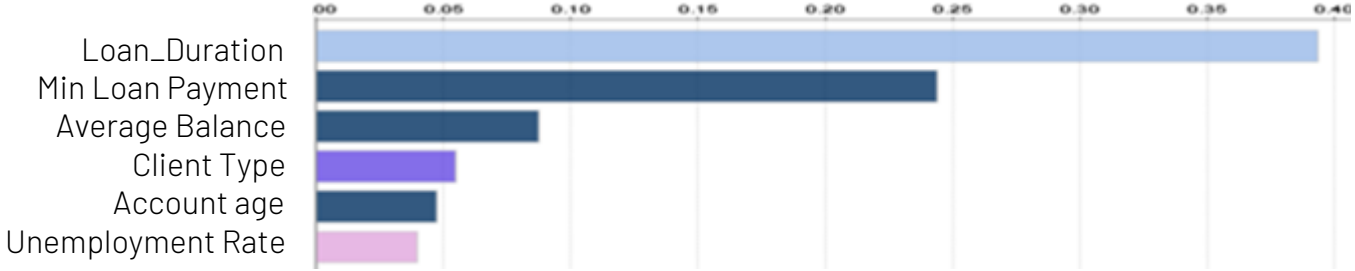
- **Reference date** 1 / Feb / 2017
- **SAP Tool** Predictive Factory

Succeeded	12/15/20 8:05 PM	78.86%	97.97%

Model Performance: Variable Contribution

Top variable contribution to predicting the defaulters:

- Loans Duration has the highest contribution to defaulters



1. Second Scenario

Enhancing Communications Strategy

YOUNG PROFESSIONAL PROGRAM

Enhancing Communications Strategy - Overview

Business Goals

The project has the following objectives:

- The profiles of customers in each group will be analyzed by marketers to develop personalized product and communication strategies for each group.
- To establish better customer relationship management strategies.
- Improve existing services and increase customer satisfaction and loyalty.

Business Success Criteria

This project will be judged a success if:

- Cross-sales increase by 5%
- Customer click-through rate on promotional offers increases from 5% to 7%

Data Science Goals

This project has the following data science goals:

- Create a supervised clustering model that strategically segments the customer base. The clusters will be developed based on customer data, demographic variables, geographic locations, transactional and product history.
- The target is estimated earnings per customer.

Data Science Success Criteria

The following is a summary of steps that you can take to validate a clustering model in SAP Predictive **Analytics:**

- Clustering results must have business sense. This is verified by analyzing the profiles for each of the clusters.
- Check if cluster sizes are relatively balanced. If there are clusters with high frequency, then this might indicate there are not enough clusters. If there are clusters with very low frequency, then this might indicate there are too many clusters.
- In this scenario, the number of segments should be greater than 3, but less than 8 ($3<k<8$).
- Check Predictive Power (KI) and Prediction Confidence (KR): KI should be high (>0.6) and KR should be greater than 0.95, especially if you want to use the model to make predictions (i.e. assign new customer to groups).
- Evaluate if there are inconsistencies in the categorical variables in each cluster by checking that the category profile has business sense.
- Check the profile of continuous variables. They should have business sense and in most cases be continuous within a cluster. For example, age should be a range and not a mix of ranges and distinct values.
- The model will be operationalized and applied monthly on the 1st day of the month, using Predictive Factory.

Model Performance: KI & KR

kc_TARGET		
Predictive Power (KI)	**0.6570**	1
Prediction Confidence (KR)	**0.9766**	2

TARGET		
Initial Number of Clusters	**6**	
Final Number of Clusters	**6**	
Overlap	**27.21%**	3
Percentage of Unassigned Records	**1.58%**	4

The winner model

Model	Winner	KI	KR	Initial Number of Clusters	Final Number of Clusters	Overlap	Percentage of Unassigned Records
Engine0	false	0.5725	0.9866	4	4	12.36%	7.12%
Engine1	false	0.6147	0.9906	5	5	24.55%	5.16%
Engine2	true	0.6570	0.9766	6	6	27.21%	1.58%
Engine3	false	0.6197	0.9911	7	7	26.36%	3.35%

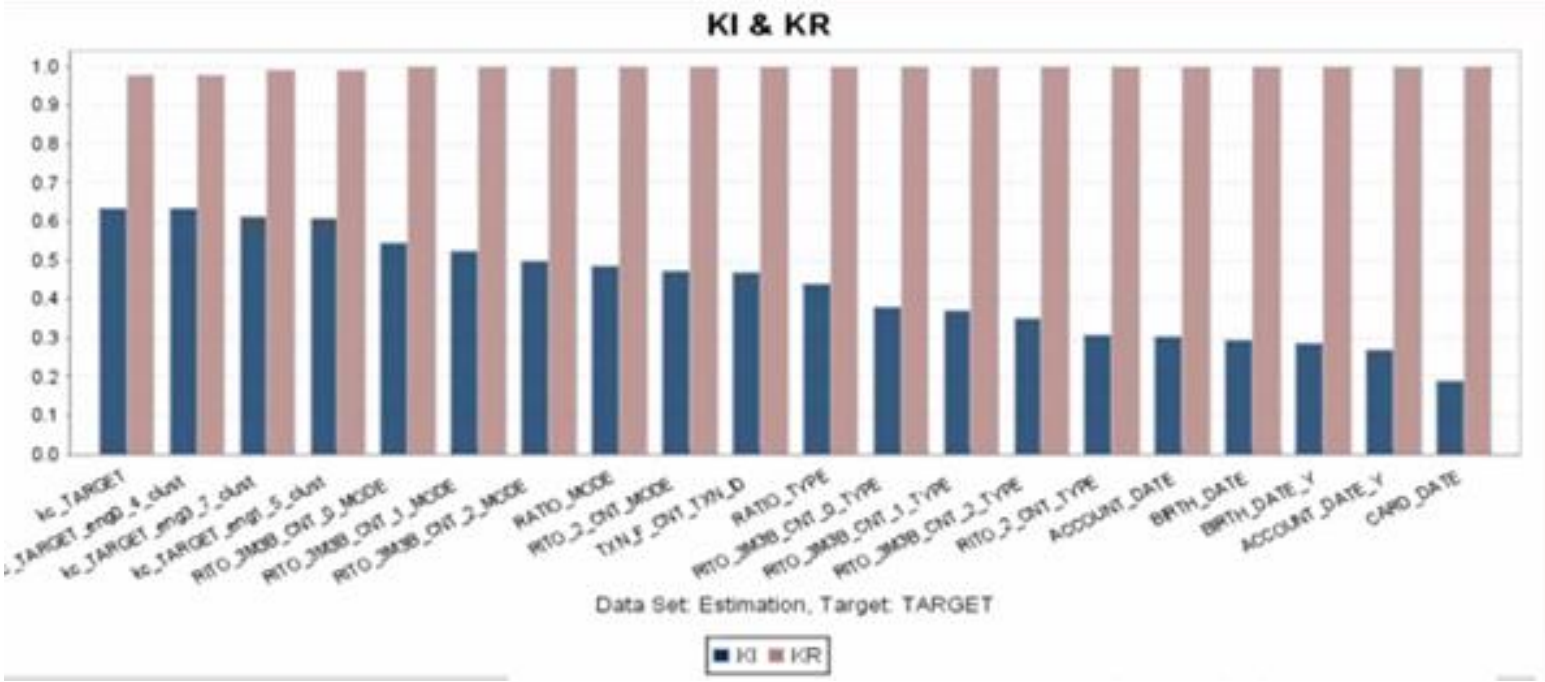
1

- We Used 4 to 7 Cluster
- The model with 6 clusters was choosing as Winner .
- **The Winner model**

  The less percentage of unassigned Records.

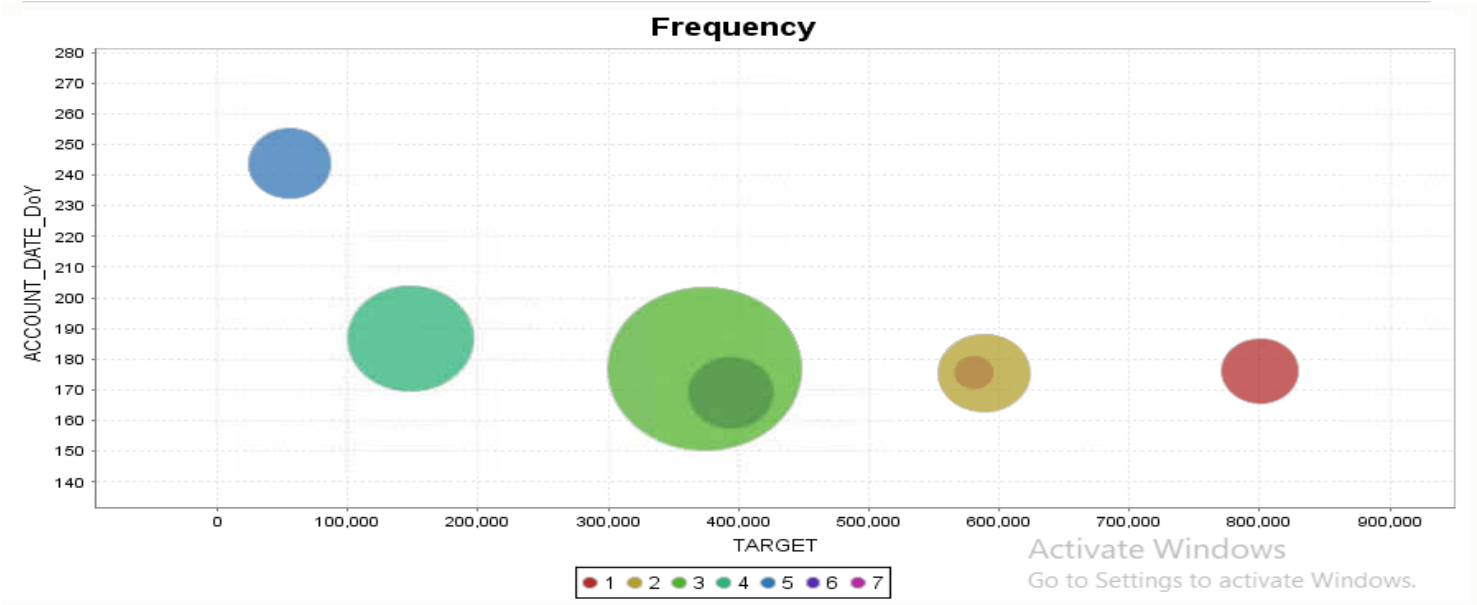
  Have 27.21% percentage overlap Records.

KI & KR for Each Variables:



Bubble Chart for the clusters

- Cluster 3 (light green) is the largest number of assigned records



Relative Target Means:

Clusters contribution to the target:

Target Mean & Standard Deviation of clusters

As we can see here the most interesting cluster is **cluster 1** with highest target mean and the highest standard deviation which means that cluster 1 had the highest average of income .

Cluster	Frequency	Target Mean	Target Standard Deviation
1	9.63%	800,200	460,793
2	12.84%	588,222	338,210
3	34.02%	374,270	236,251
4	19.77%	148,301	138,148
5	10.91%	55,367.8	61,802.5
6	11.25%	394,215	379,135

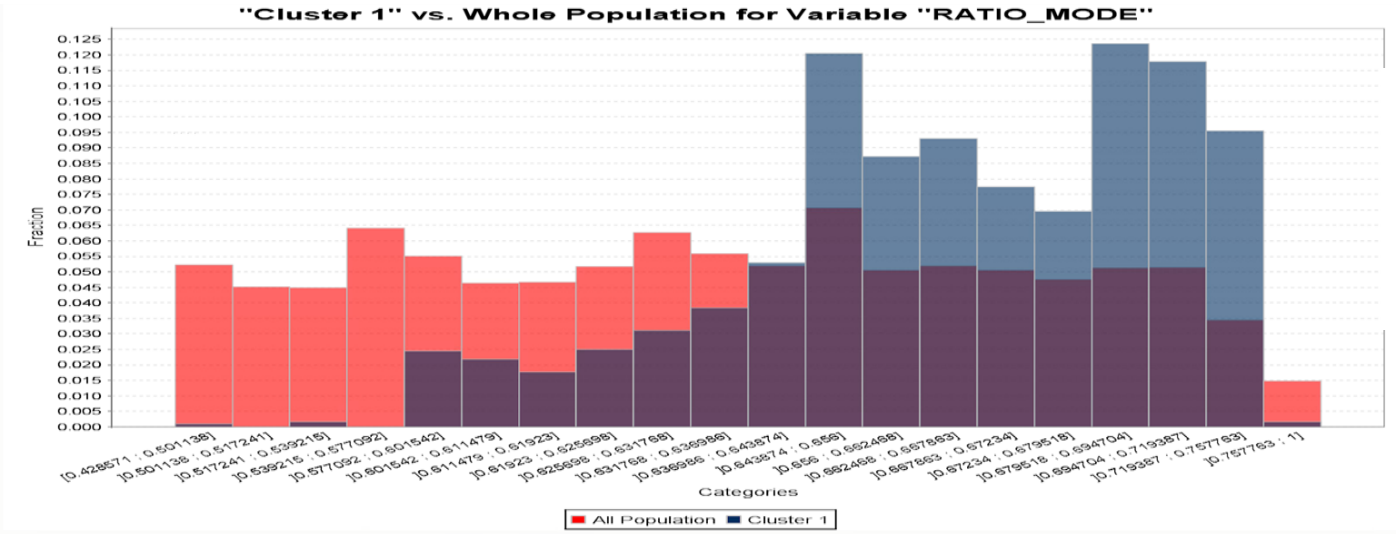
Checking For Missing Values in Target

Variable	Value	Storage	Missing Value Count	Missing Value Weight Role
KxId	nominal	integer		skip
KxTimeStamp	nominal	date		skip
TARGET	continuous	number	0	0 target

Here we're see that your target variable didn't include any missing values.

Cluster Profiles:

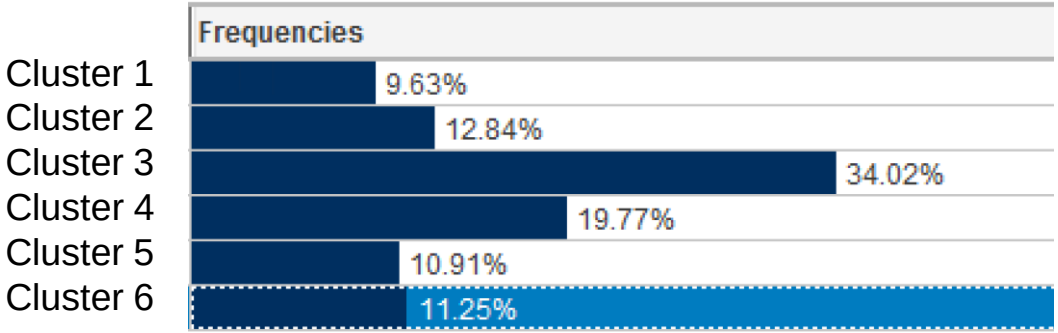
The variabel that have the highest influence on the cluster was the ratio of mode which means that when you targeting to communicate with your customers you should targeting them based on the mode of the transactions .



KL	0.41
KL Significance	1
Overall mean:	0.628408
Cluster mean:	0.66929

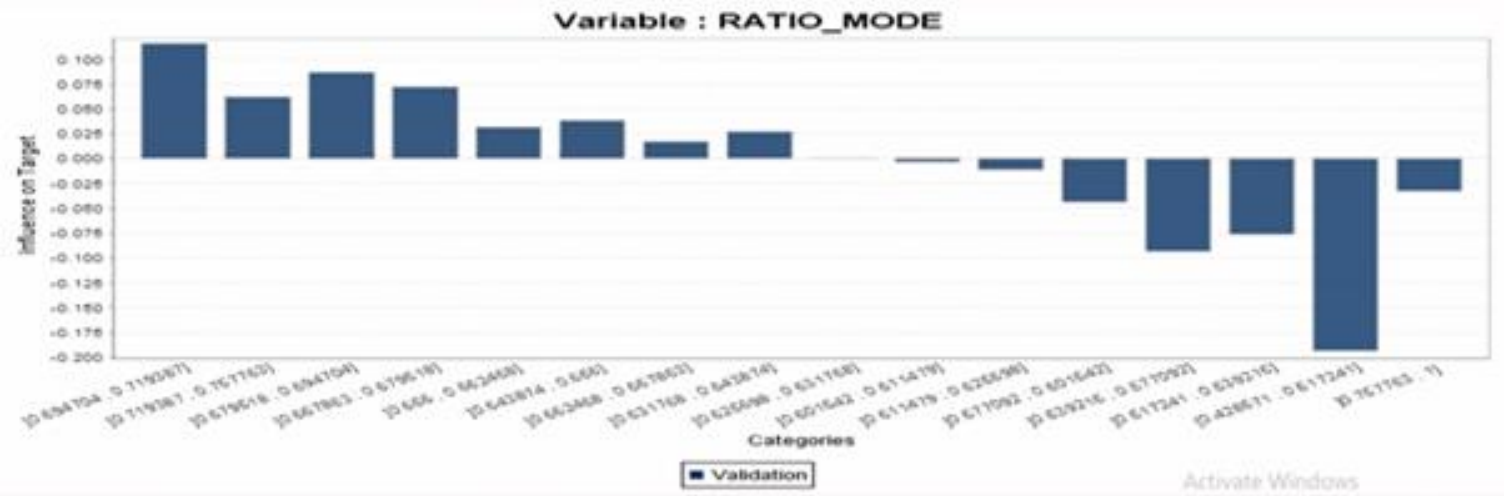
Clusters Frequency:

The amount of frequency of records in each cluster:



Contribution :

The variable ratio of mode have positive contribution on target variable from 63% - 71%
And from 62% and less it have negative contribution on the target .



Predictive factory - Deploy

BOC_04

Settings | Variable Metadata | Models | Tasks | Variable Statistics

Models (4)

Search | Import

Name	Type	Creation Date	Author
Cluster_TSP_SC2Population	Clustering	December 15, 2020 7:04 PM	STUDENT05
Cluster_TSP_SC2Population_2	Clustering	December 15, 2020 7:25 PM	STUDENT05
TARGET_DEPOSIT_NEXT_3_MONTHS_Sc3Population	Regression	December 15, 2020 5:01 PM	STUDENT05
TARGET_INH_SPopulation	Clustering	December 15, 2020 5:32 PM	STUDENT05

Deploying the Cluster model in predictive factory

Predictive factory - model information

Cluster_TSP_SC2Population_2
Active:Version 1

Versions Tasks Monitoring

Model Versions (1)

Publish Set as Active

Name	Training	Reference Date	Number of Clusters	Variable Count	Record Count	
Version 1 Imported from source 'STUDENT05'	Succeeded	January 1, 2017 12:00 AM	6	55	26845	Active

1. Third Scenario

Improving Customer Satisfaction

YOUNG PROFESSIONAL PROGRAM

Improving Customer Satisfaction - Overview

Business Goals

This project has the following objectives:

- Improve customer loyalty and satisfaction

Business Success Criteria

This project will be judged a success if:

- Customer attrition rate decreases from 17% to 10%
- Customer satisfaction increases from 75% to 85%

Data Science Goals

- Create a **regression model** that estimates the deposit sum for the next three months following a latency period.
- **Latency period:**
  - 1 month
- **History period:**
  - 6 months
- **Target Period:**
  - 3 months
- Population filters: exclude new customers who joined the Bank less than six months ago
- Create a customer profile of high value customers
- Model will be operationalized and applied monthly using Predictive Factory

Data Science Success Criteria

- Predictive Power: > 0.7
- Prediction Confidence: > 0.95
- Model performance in evaluation period corresponds to performance in training period
- Model will be operationalized and applied monthly on the 1st day of the month, using Predictive Factory

Model Overview

Distribution of dataset for regression

Data Set	Number of Records
Estimation	18,255
Validation	6,236

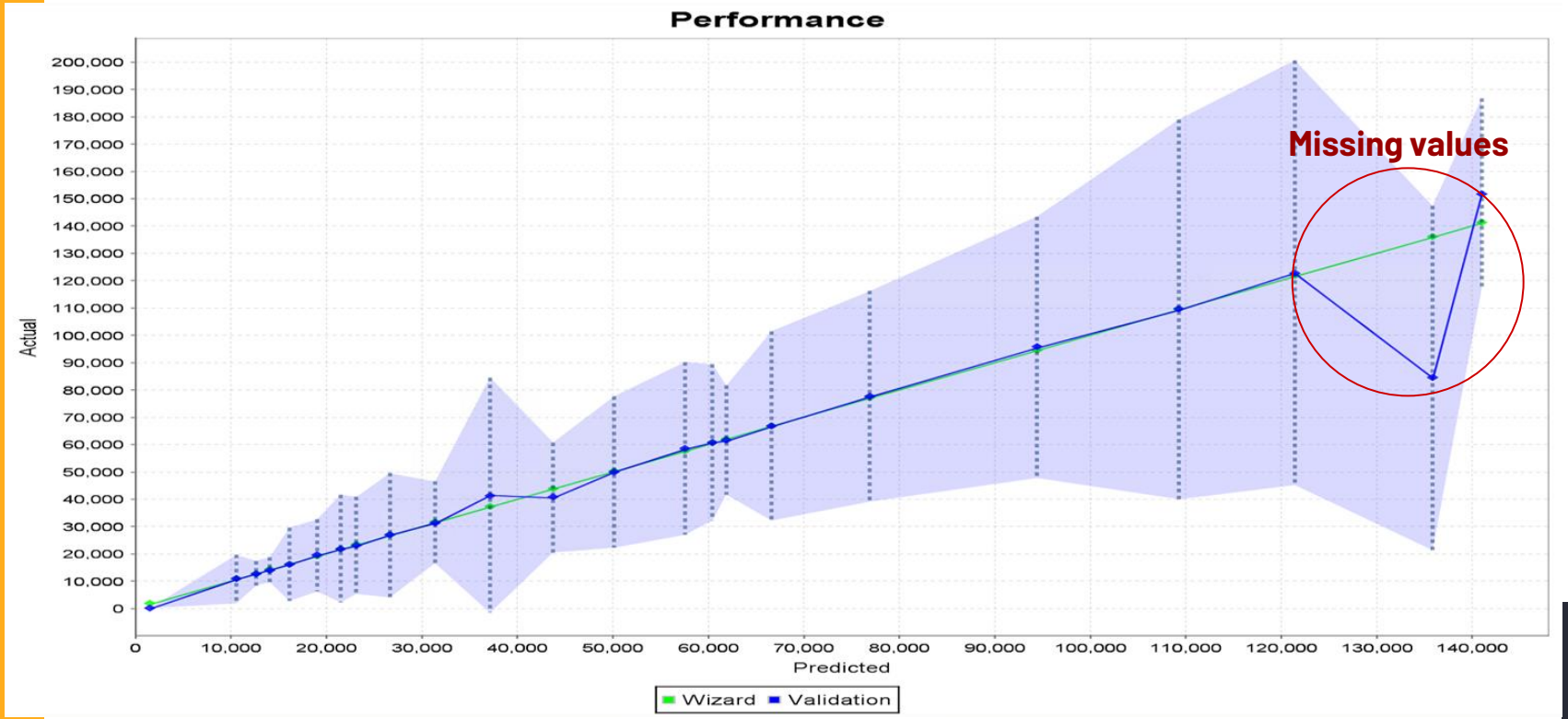
Continuous Targets (Number)

TARGET_DEPOSIT_NEXT_3_MONTHS2	
Min	0
Max	399,186
Mean	51,234.1
Standard Deviation	48,947.9

Selection Process Selected Iteration

2	
Predictive Power (KI)	0.8145
Prediction Confidence (KR)	0.9938
Nb. Variables Kept	8

Model performance - Regression Model

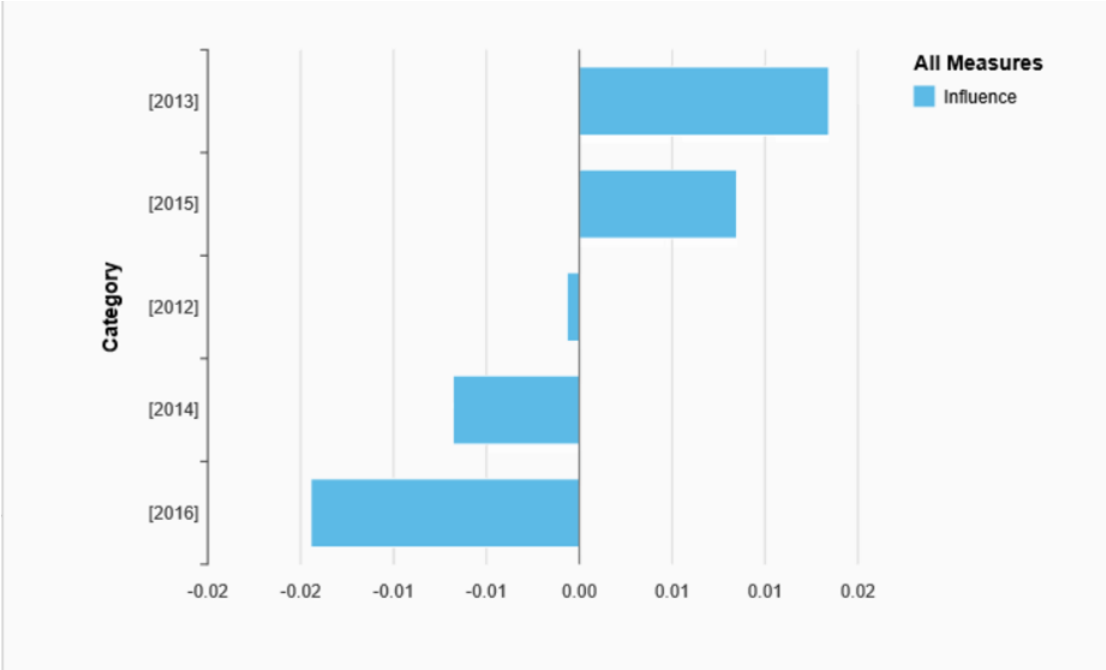


Variables Contribution

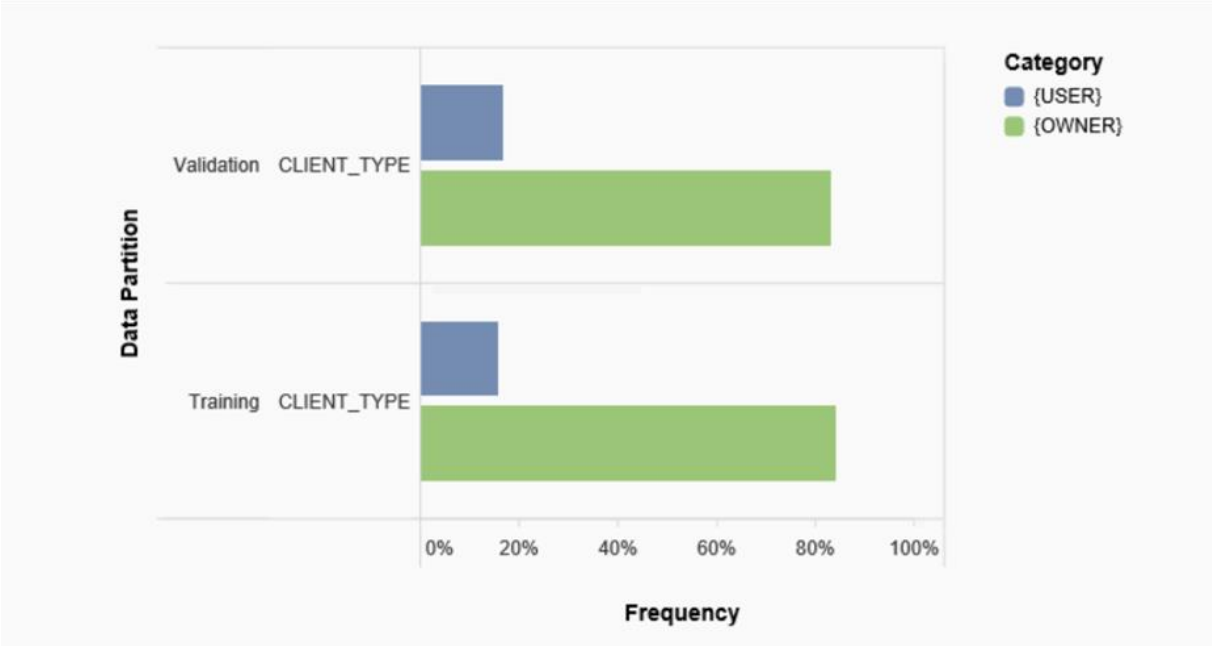
Variable Contributions

Variable	Contribution
MAX_BALANCE_1M6B_MAX_0_BALANCE	81.36%
MIN_BALANCE_1M6B_MIN_0_BALANCE	18.64%

Created accounts in Years - influance

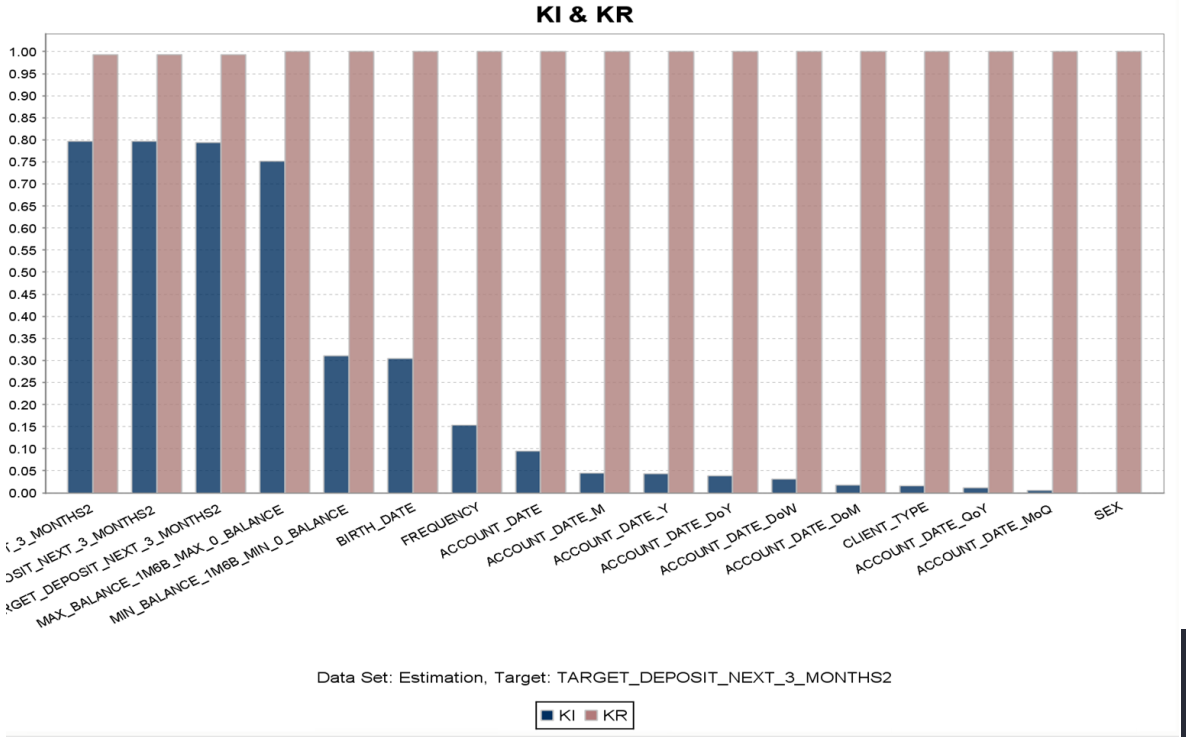


Type of Clients



Performance of KI & KR

KI & KR for each selected variable



Performance Indicators

- R-Squared has a low value due to unexplainable error after net value of 120K

Indicator		Estimation	Validation
L1	(Manhattan)	18,237.2	18,982.8
L2	(Euclidian)	34,123.3	35,120.1
LInf	(L norm (infinity))	309,194	308,681
ErrorMean	(Mean of error)	259.741	159.887
ErrorStdDev	(Error Standard deviation)	34,122.3	35,119.8
R2	(R Squared)	0.514	0.506

Predictive factory - Deploy

BOC_04

Settings | Variable Metadata | Models | Tasks | Variable Statistics

Models (4)

Search | Import

Name	Type	Creation Date	Author
Cluster_TSP_SC2Population	Clustering	December 15, 2020 7:04 PM	STUDENT05
Cluster_TSP_SC2Population_2	Clustering	December 15, 2020 7:25 PM	STUDENT05
TARGET_DEPOSIT_NEXT_3_MONTHS_Sc3Population	Regression	December 15, 2020 5:01 PM	STUDENT05
TARGET_INH_SPopulation	Clustering	December 15, 2020 5:32 PM	STUDENT05

Performance Indicators - Predictive factory

TARGET_DEPOSIT_NEXT_3_MONTHS_Sc3Population

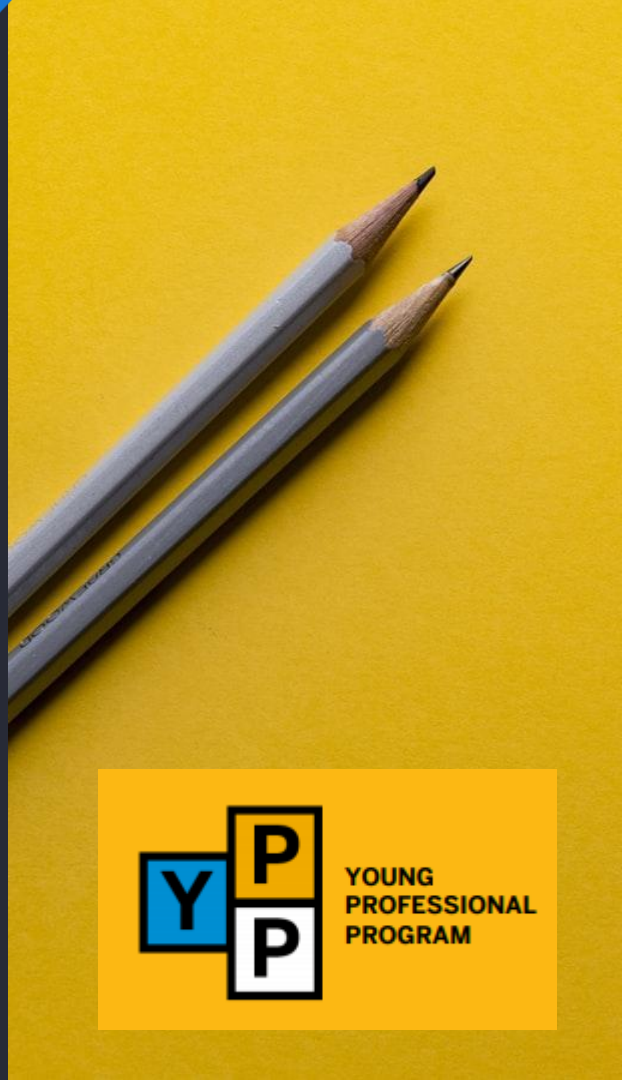
Active:Version 2

Versions | Tasks | Monitoring

Model Versions (2)

Publish | Set as Active | Duplicate

Name	Training	Reference Date	Predictive Power	Prediction Confidence	Variable Count	Record Count	
Version 2 Imported from source 'STUDENT05'	Succeeded	January 1, 2017 12:00 AM	78.88%	99.29%	2	23986	Active
Version 1 Imported from source 'STUDENT05'	Succeeded	January 1, 2017 12:00 AM	78.88%	99.29%	2	23986	



YOUNG PROFESSIONAL PROGRAM

Thanks!

You can find us at:

- Feryal Almutairi
- Maisoun Alshahrani
- Turki Alsulaimani
- Noura Alharbi
- Saad Alraozuq
- Rahaf Alawwad

Extra (Technical)

Task Application Attempts: 1

APPLY_YHE_MODEL_1

Settings Runs

Task Runs (7) Run Task Now

Run	Status	Reference Date	Start Date	End Date
6	Succeeded	November 1, 2017 8:18 PM	December 15, 2020 8:18 PM	December 15, 2020 8:19 PM
5	Warning	October 1, 2017 8:16 PM	December 15, 2020 8:16 PM	December 15, 2020 8:17 PM
4	Warning	November 1, 2017 8:14 PM	December 15, 2020 8:14 PM	December 15, 2020 8:15 PM
3	Warning	November 1, 2017 8:13 PM	December 15, 2020 8:13 PM	December 15, 2020 8:14 PM
2	Warning	November 1, 2017 8:11 PM	December 15, 2020 8:11 PM	December 15, 2020 8:11 PM

Target Statistics: 1

Target Statistics

Data Partition	Target Category	Frequency
Training	0	90.73%
Training	1	9.27%
Validation	0	91.01%
Validation	1	8.99%

Missing Value substitution (attempt): 1

```
  case
   when ((("KXTempT1"."LOAN_STATUS" = N'B')
   or ("KXTempT1"."LOAN_STATUS" = N'D'))) then 1
   else 0
  end as "target",
  case
   when ((("KXTempT1"."LOAN_DURATION" IS NULL )
   and ("KXTempT1"."LOAN_AMOUNT" <= 119000))) then 12
   when ((("KXTempT1"."LOAN_DURATION" IS NULL )
   and ("KXTempT1"."LOAN_AMOUNT" <= 236328))) then 24
   when ((("KXTempT1"."LOAN_DURATION" = cast(NULL as integer))
   and ("KXTempT1"."LOAN_AMOUNT" <= 355328))) then 36
   when ((("KXTempT1"."LOAN_DURATION" IS NULL )
   and ("KXTempT1"."LOAN_AMOUNT" <= 474328))) then 48
   when ((("KXTempT1"."LOAN_DURATION" IS NULL )
   and ("KXTempT1"."LOAN_AMOUNT" >= 590820))) then 60
   else "KXTempT1"."LOAN_DURATION"
  end as "loan_duration_modiefied"
from
 "STUDENT04"."KX_16079350211704268TemporaryStoreForUI" "KXTempT1"
```

Model Performance: Performance Metrics: 2

Sum of Squares.
Frequency.
Within Cluster Variance.
Target Mean.
Target Standard Deviation.

Clusters Estimation: 2

Target TARGET

Data Set Estimation

Cluster Id	Sum of squares	Within Cluster Variance	Frequency	Target Mean	Target Standard Deviation
1	263,270,000,000,000	136,693,000,000	9.63%	800,200	460,793
2	181,016,000,000,000	70,461,800,000	12.84%	588,222	338,210
3	178,695,000,000,000	26,247,700,000	34.02%	374,270	236,251
4	137,970,000,000,000	34,876,200,000	19.77%	148,301	138,148
5	127,119,000,000,000	58,258,200,000	10.91%	55,367.8	61,802.5
6	179,352,000,000,000	79,641,100,000	11.25%	394,215	379,135

Cluster Validation: 2

Target TARGET

Data Set Validation

Cluster Id	Sum of squares	Within Cluster Variance	Frequency	Target Mean	Target Standard Deviation
1	88,547,300,000,000	131,963,000,000	9.82%	791,477	468,495
2	64,668,000,000,000	72,906,500,000	12.98%	606,400	342,553
3	61,300,500,000,000	26,957,100,000	33.27%	373,480	232,902
4	51,285,800,000,000	36,270,000,000	20.68%	153,161	157,170
5	45,298,100,000,000	58,449,200,000	11.34%	53,580.4	58,723.5
6	54,052,800,000,000	74,452,900,000	10.62%	367,022	343,641

Task Application :3

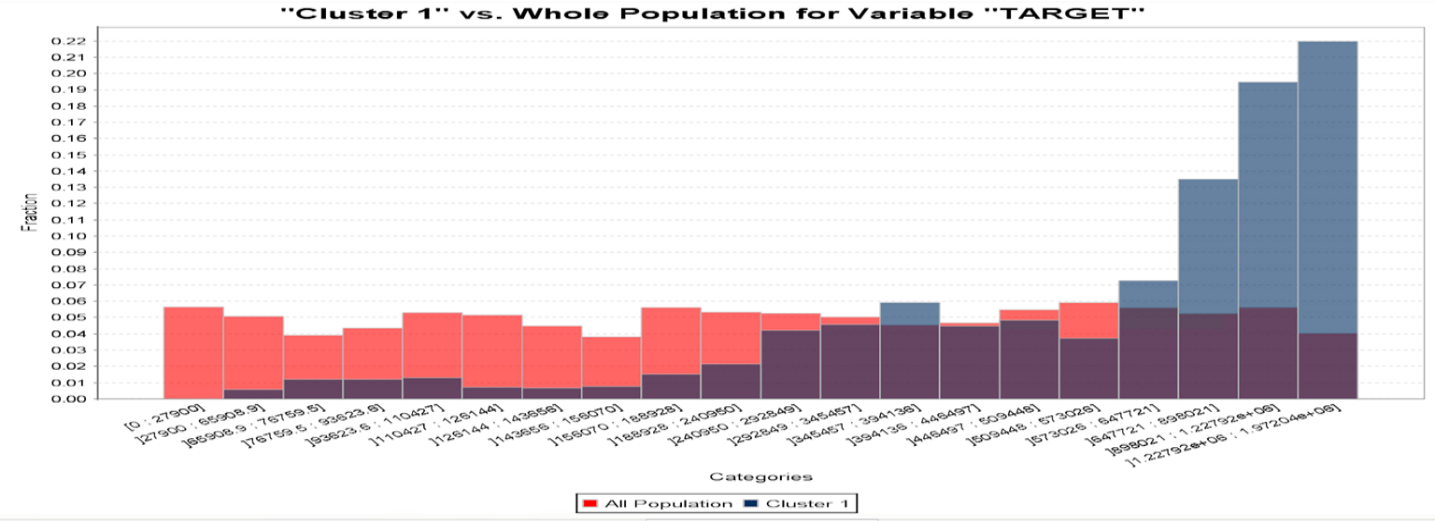
APPLY Regression MODEL

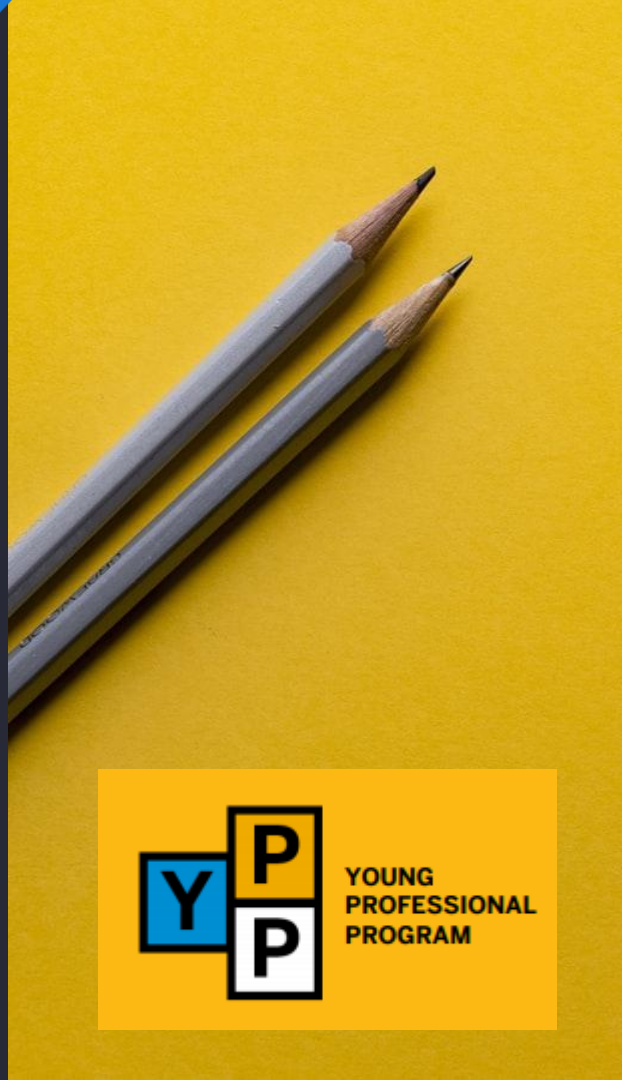
Settings Runs

Task Runs (2)

Run	Status	Reference Date	Start Date	End Date
2	Succeeded	February 1, 2017 5:17 PM	December 15, 2020 5:17 PM	December 15, 2020 5:17 PM
1	Succeeded	December 15, 2020 5:14 PM	December 15, 2020 5:14 PM	December 15, 2020 5:15 PM

Cluster Profiles: 2





The End

You can find us at:

- Feryal Almutairi
- Maisoun Alshahrani
- Turki Alsulaimani
- Noura Alharbi
- Saad Alraozuq
- Rahaf Alawwad