BANK CHURN PREDICTION & PROFILING

PRESENTED BY

CHAIWAT PREMRIDIKUL

6510414004

PROJECT'S AGENDA

00

OVERVIEW

Data flow & Dataset

01

PART 1

Data preparation and customer segmentation

02

PART 2

Defined focus group and prediction model

03

PART 3

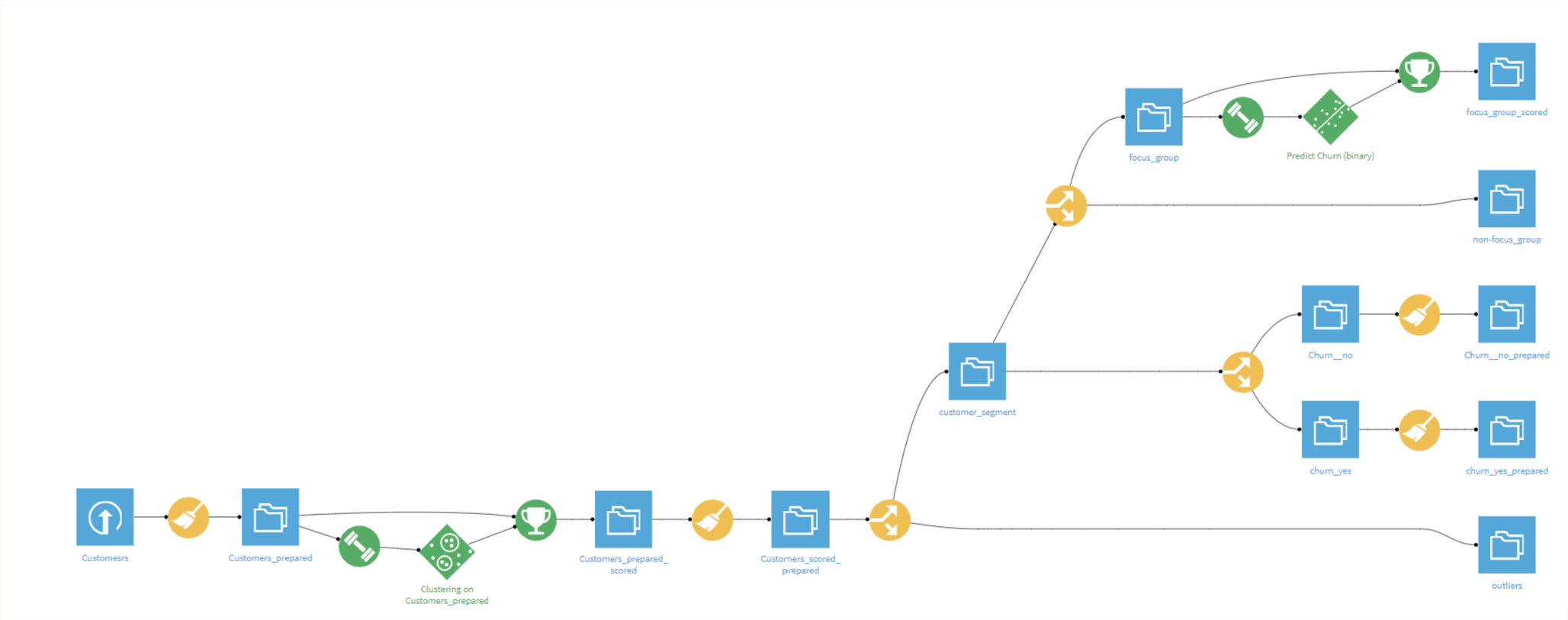
Customers and Chureners Profiling

05

PART 5

Business Impact

DATA FLOW


Customesrs
Customers_prepared
Clustering on Customers_prepared
Customers_prepared_ scored
Customers_scored_ prepared
customer_segment
focus_group
Predict Churn (binary)
focus_group_scored
non-focus_group
Churn__no
Churn__no_prepared
churn_yes
churn_yes_prepared
outliers

DATASET

Bank churn customers dataset contained these 13 schema

CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exi
string	string	string	string	string	int	int	float	int	boolean	boolean	float	bool
Integer	Text	Integer	Country	Gender	Integer	Integer	Decimal	Integer	Boolean	Boolean	Decimal	Boo
15634602	Hargrave	619	France	Female	42	2	0.0	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0.0	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
15592531	Bartlett	822	France	Male	50	7	0.0	2	1	1	10062.8	0
15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
15737173	Andrews	497	Spain	Male	24	3	0.0	2	1	0	76390.01	0
15632264	Kay	476	France	Female	34	10	0.0	2	1	0	26260.98	0
15691483	Chin	549	France	Female	25	5	0.0	2	0	0	190857.79	0
15600882	Scott	635	Spain	Female	35	7	0.0	2	1	1	65951.65	0
15643966	Goforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0

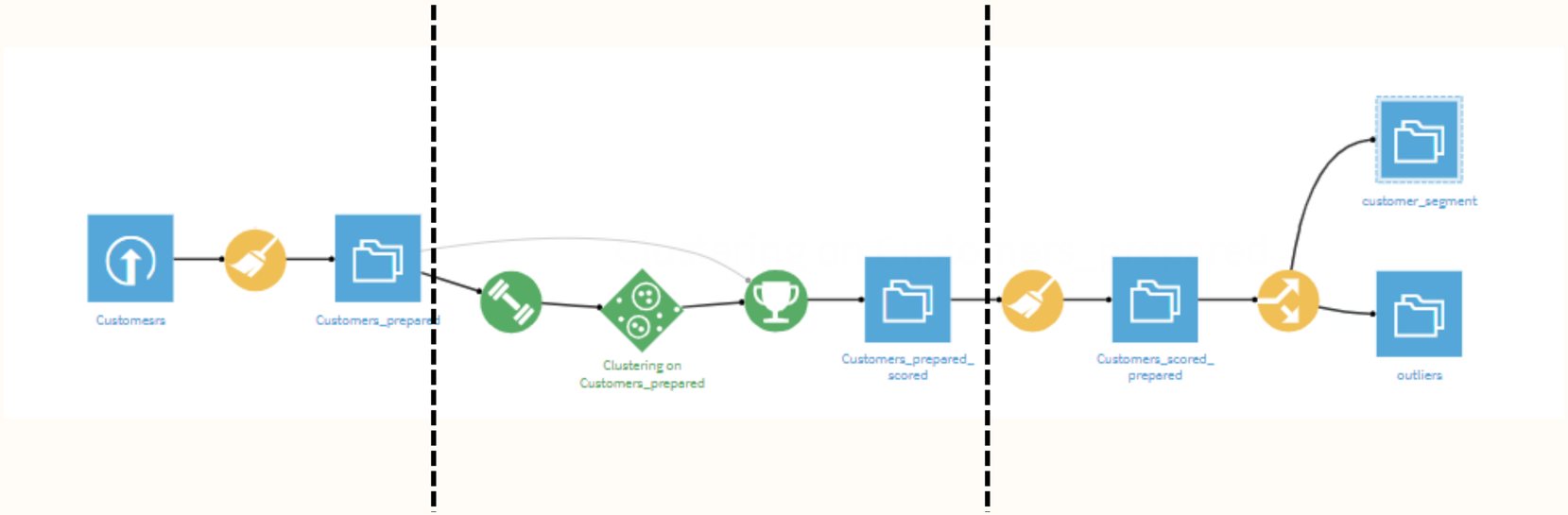
COLUMN COUNT	FILE COUNT	SIZE	RECORD COUNT
13	1	621.05 KB	10 000
2023-06-27 15:57	2023-06-27 15:57	2023-06-27 15:57	2023-06-27 16:07

Schema (13)

Search

CustomerId	string
Surname	string
CreditScore	string
Geography	string
Gender	string
Age	int
Tenure	int
Balance	float
NumOfProducts	int
HasCrCard	boolean
IsActiveMember	boolean
EstimatedSalary	float
Exited	boolean

PRAT 1



STEP 01

Data preparation

STEP 02

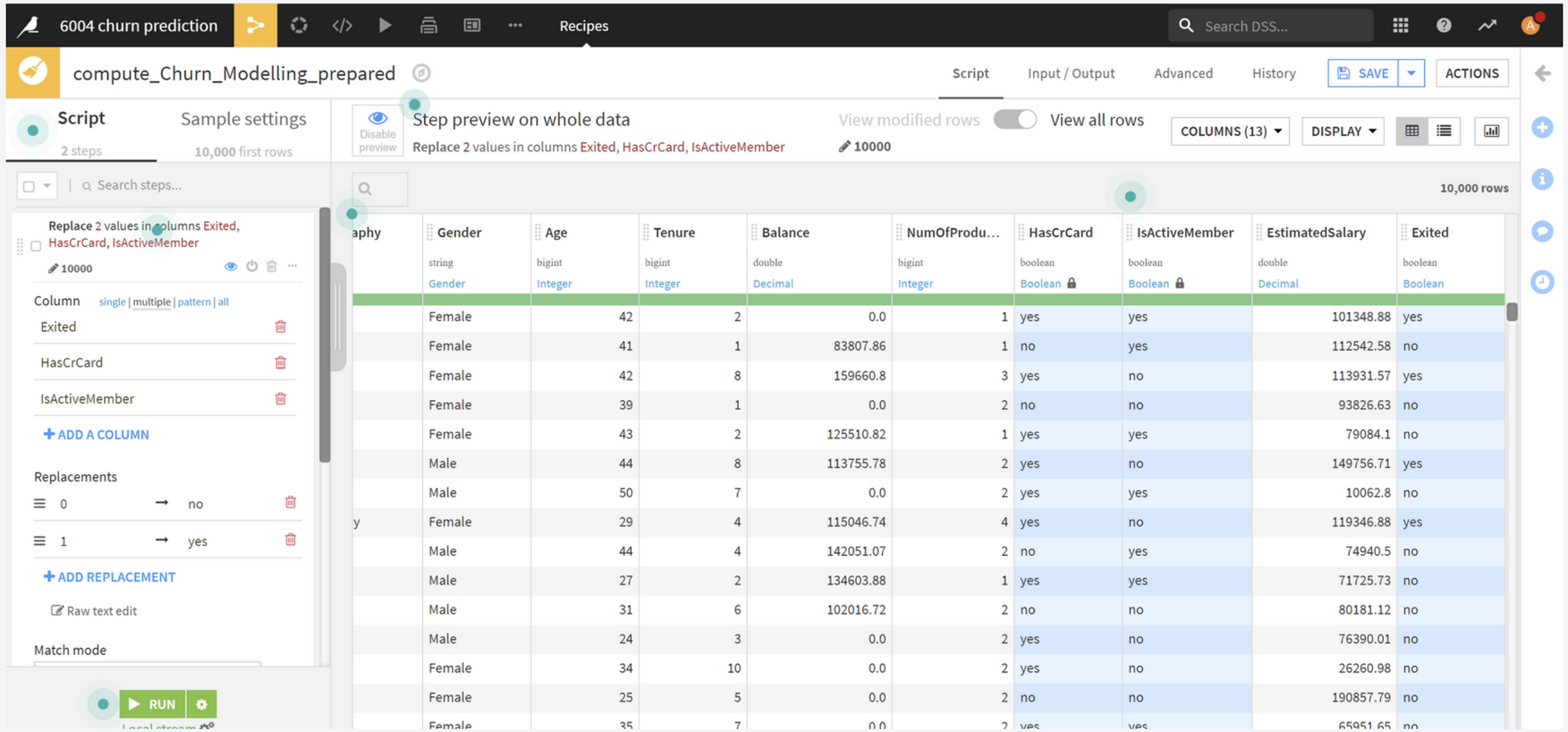
Cluster modeling

STEP 03

Separate outlier

STEP 1: DATA PREPARATION

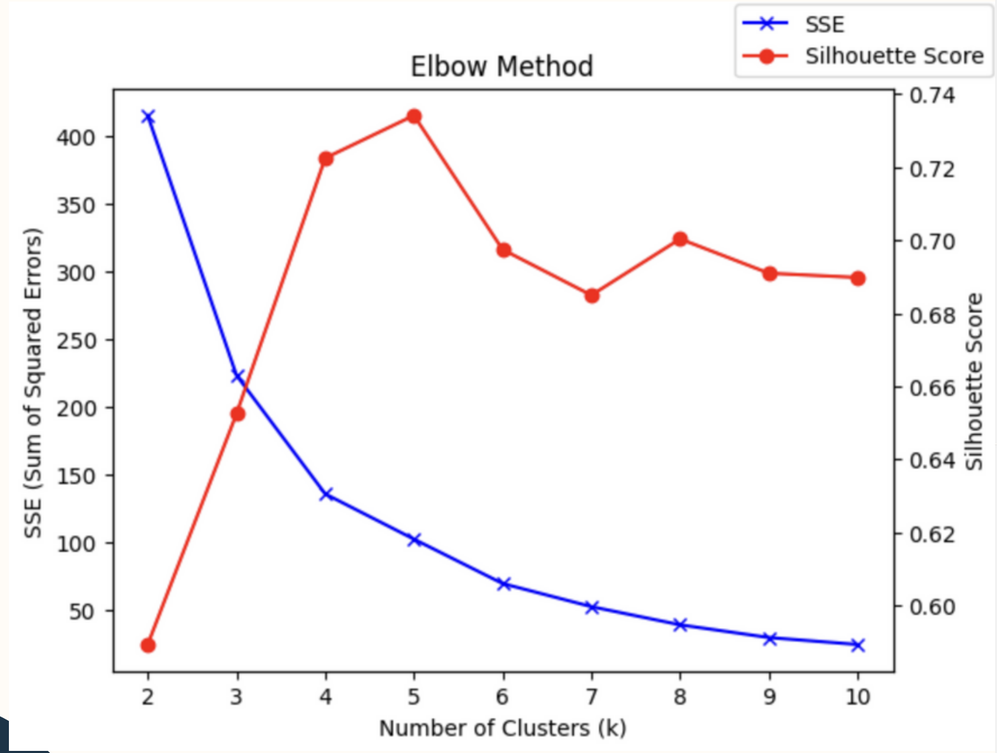
- CHANGE COLUMN TYPE
- REPLACE VALUE
- CLEAN OUTLIER



STEP 2: CLUSTER

DESIGN MODEL

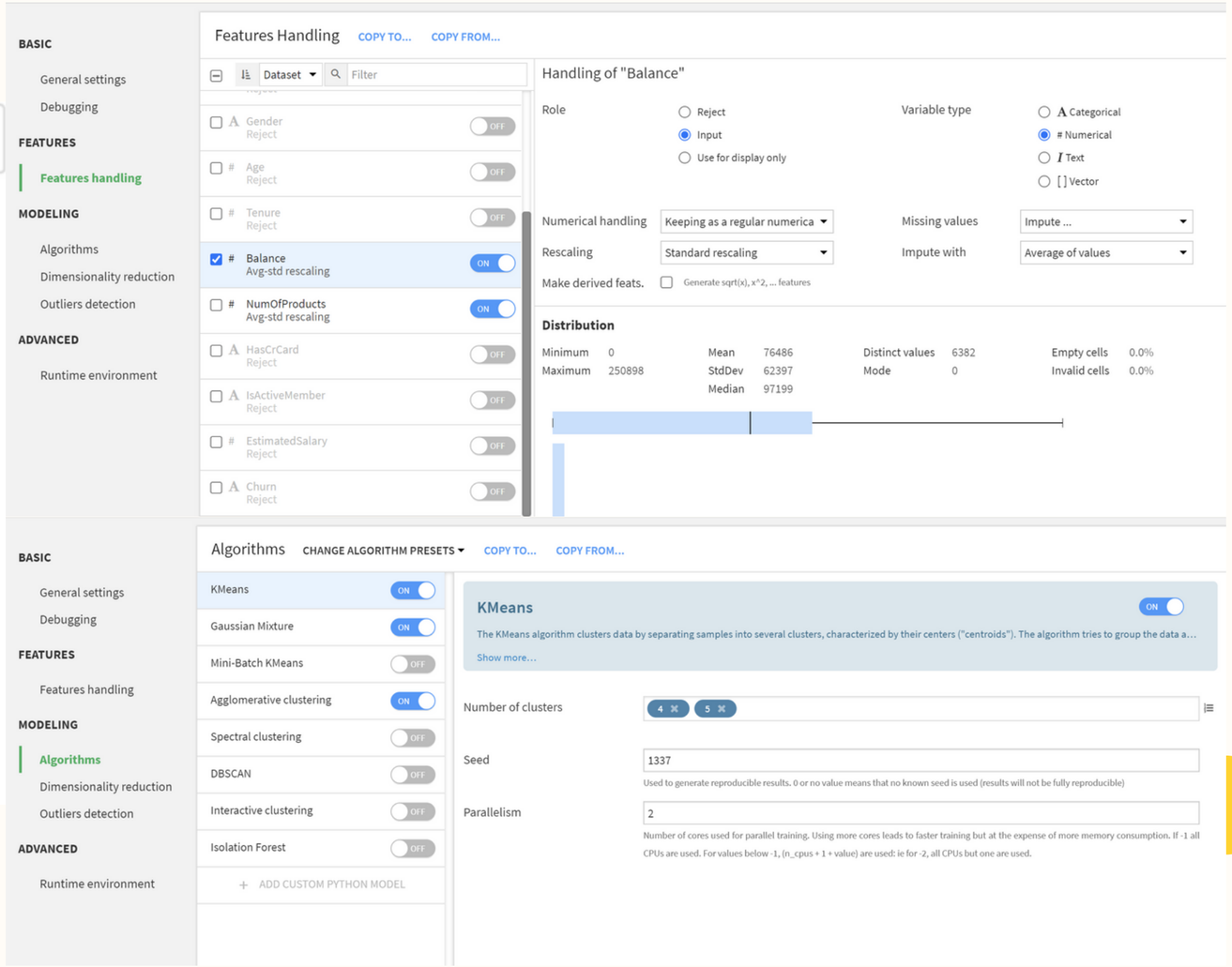
Elbow Method & Silhouette Score



to find number of K for clustering

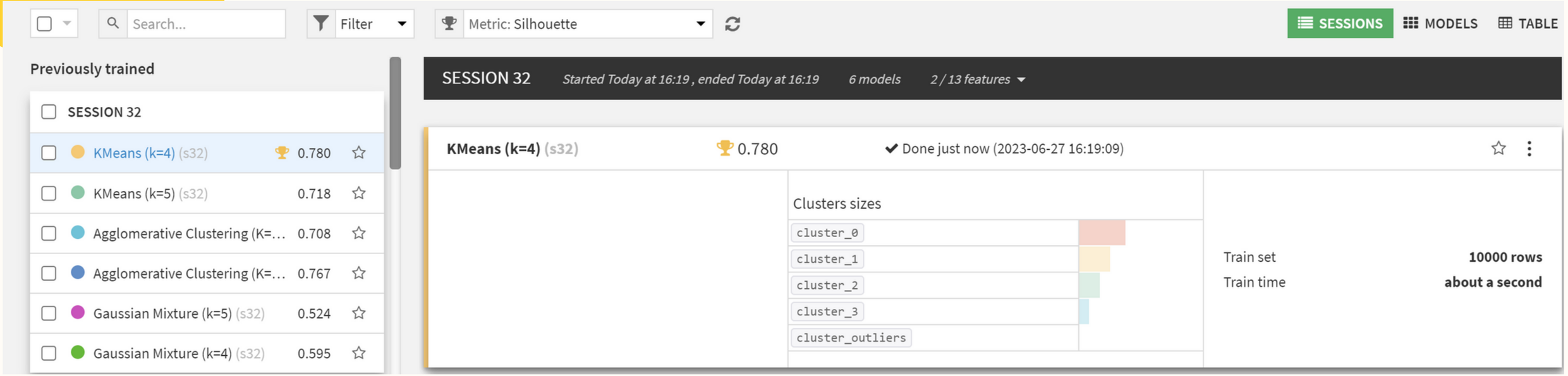
K = 4, 5

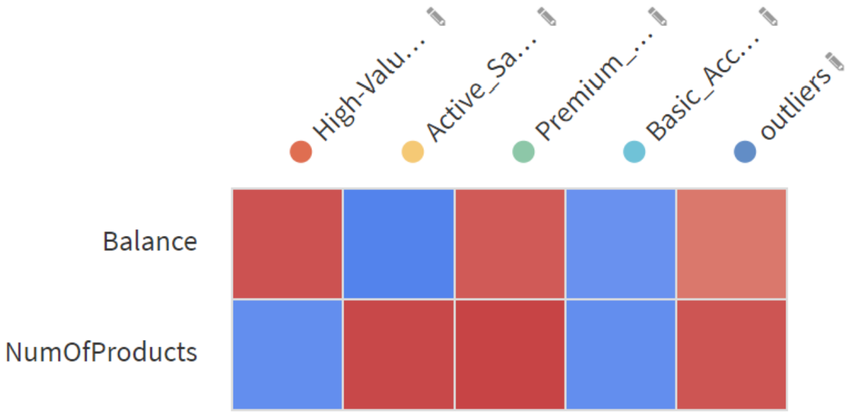
Feature Handling & algorithms



STEP 2: CLUSTER

TRAINING RESULT





01

BASIC ACCOUNT HOLDERS

- low balance
- low no. of product

02

ACTIVE SAVERS

- low balance
- high no. of product

03

HIGH-VALUE CUSTOMERS

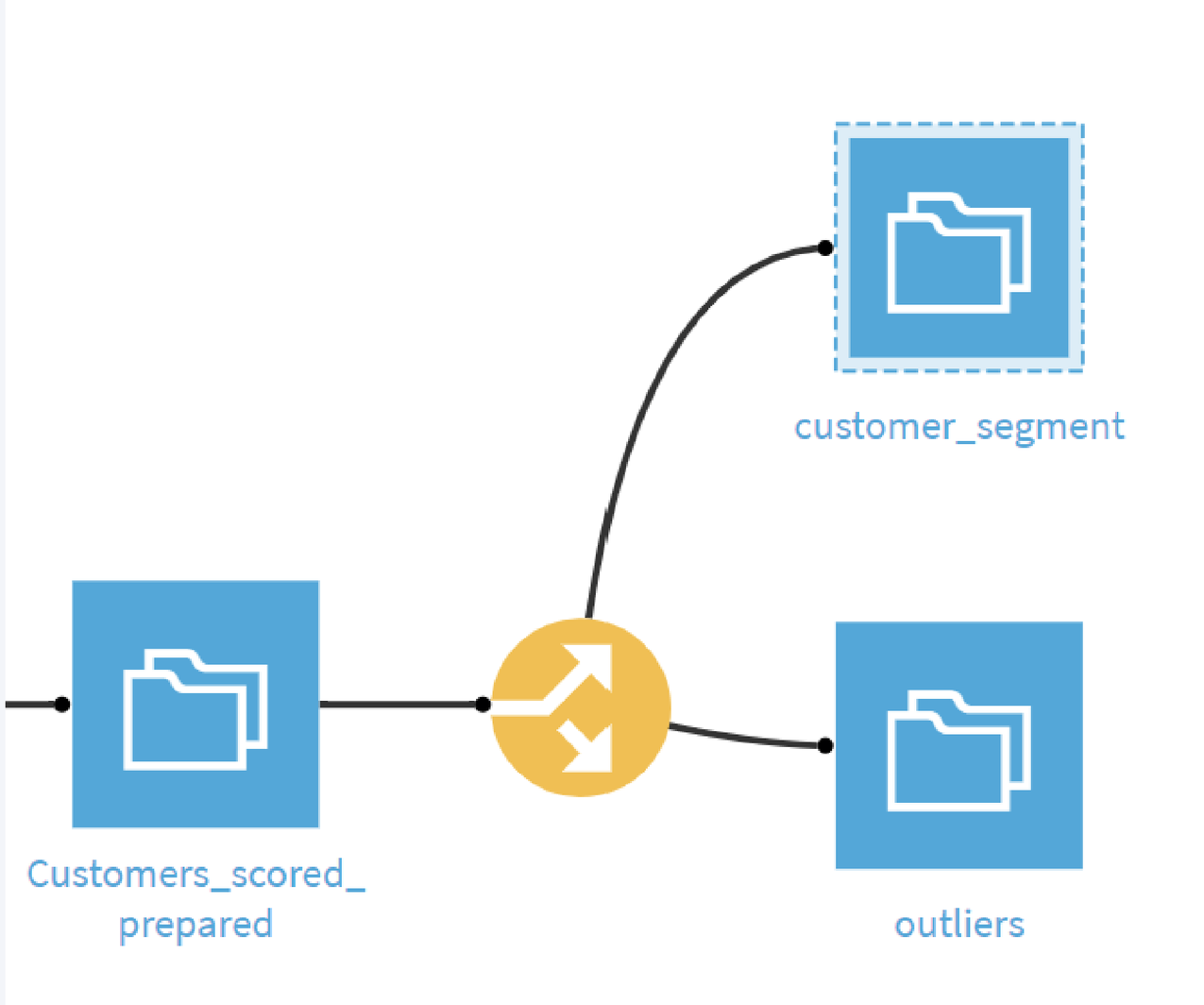
- high balance
- low no. of product

04

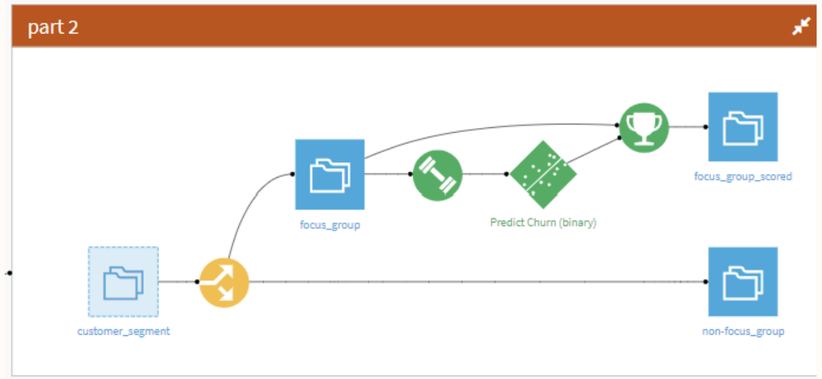
PREMIUM CLIENTS

- high balance
- high no. of product

STEP 3: SEPARATE CLUSTER OUTLIER



PART 2



DEFINED FOCUS GROUP

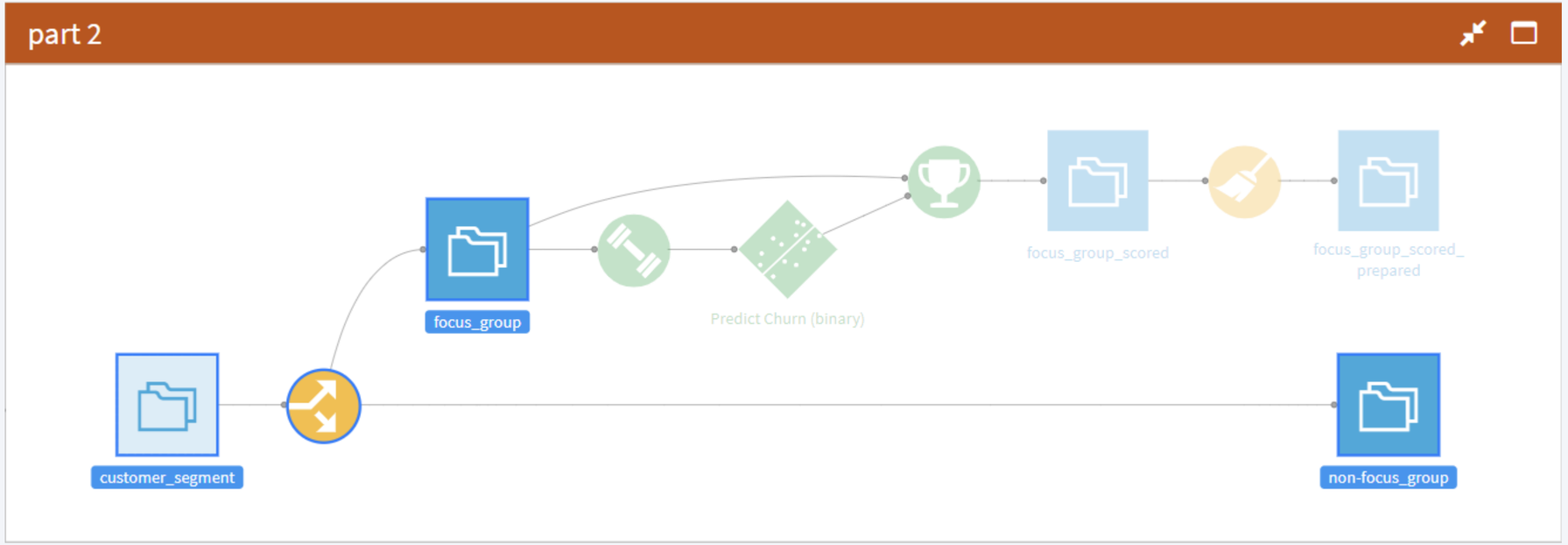
CHURN PREDICTION

FOCUS TO PREDICT CHURN IN BASIC ACCOUNT HOLDERS & HIGH-VALUE CUSTOMERS WHICH HAVE HIGH CHURN RATIO

Count by Churn and Customer_types

		Churn	
Customer_types	Value	no	yes
Active_Savers	Count of records	94%	6%
Basic_Account_Holders	Count of records	64%	36%
High-Value_Customers	Count of records	74%	26%
Premium_Clients	Count of records	87%	13%

SPLIT DATA INTO FOCUS & NON-FOCUS GROUP



DESIGN MODEL

PREDICT CLASS IMBALANCE

Target

Prediction type: Two-class classification

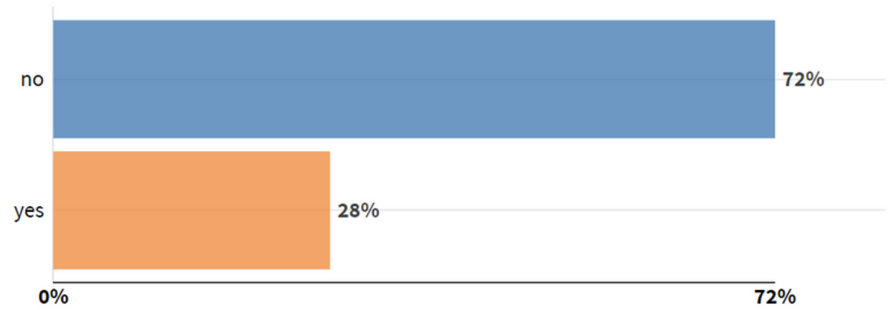
Target: Churn

Partitioned models

Partitioning: Not available: input dataset is not partitioned.

Target classes

Proportions of classes in the guess sample



RESAMPLING & K-FOLD CROSS TEST

Train / test set for final evaluation

Policy: Split the dataset

Time ordering

Enabled: OFF

Sampling & Splitting

If your dataset does not fit in your RAM, you may want to subsample the set on which splitting will be performed.

Sampling method: Class rebalance (approx. nb. recor

Nb. records: 2800

Column: Churn

Split: Randomly

For more advanced splitting, use a split recipe, and then use "Explicit extracts from two datasets" policy

K-fold cross-test: Gives error margins on metrics, but greatly increases training time

Number of folds: 5 — Number of folds to divide the dataset into

Random seed: 22422 — Using a fixed random seed allows for reproducible result

Stratified: Preserve target variable distribution within every split

Grouped: Rows with the same value for the group column are assigned to only one fold

DESIGN MODEL

METRIX EVALUATION: AUC & F1 SCORE

Metric

Hyperparameter optimization and model evaluation

Optimize model hyperparameters for: AUC — See the documentation

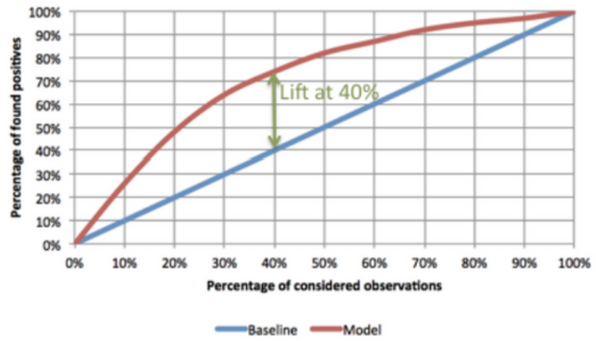
Optimize threshold for: F1 Score — See the documentation

Evaluate Model on: + NEW CUSTOM METRIC — See the documentation

Lift

The cumulative lift of a model is the additional proportion of positive records found by a model on a given proportion of the test set, compared to a random model.
DSS computes the cumulative lift for a given proportion that you need to specify.

Compute lift at: 40 %



Cost matrix

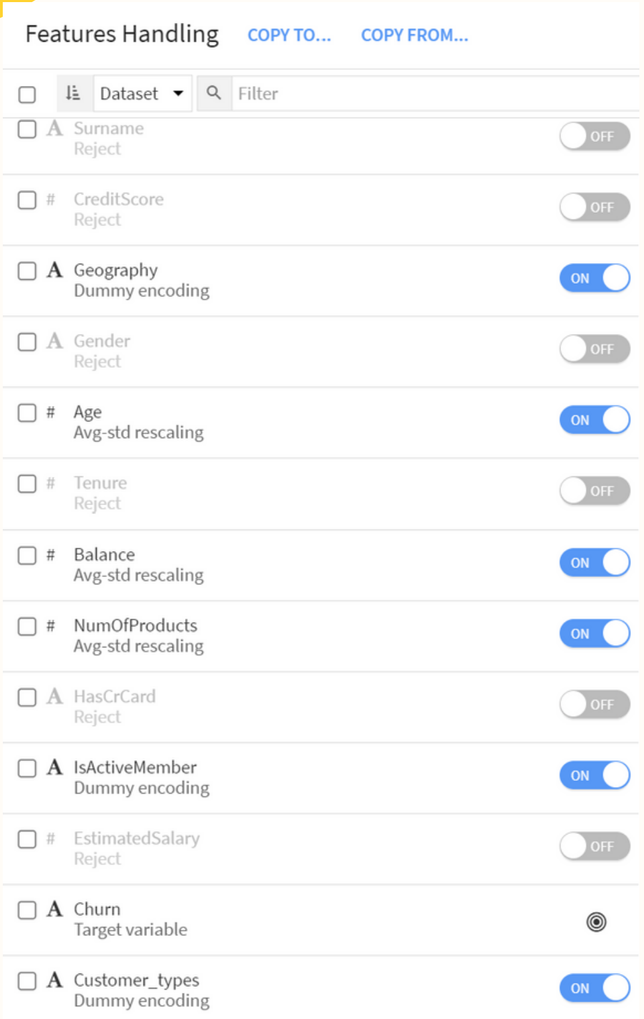
The cost matrix evaluates a "gain" brought by this model based on the following hypothesis

If the model predicts that **Churn** is true
- and it is indeed true, the gain is: -3
- but it is not true, the gain is: -3

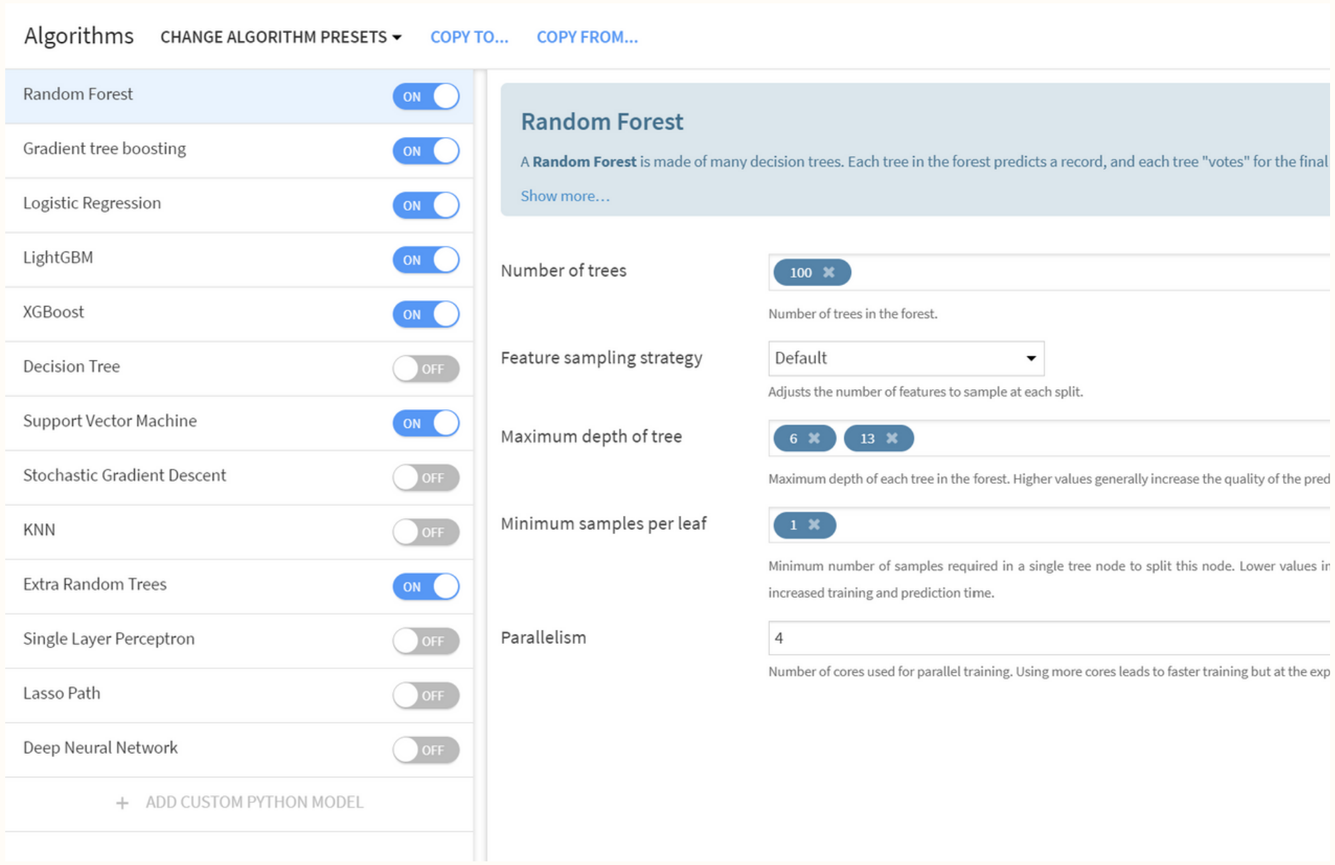
If the model predicts that **Churn** is false
- and it is indeed false the gain is: 0
- but it is actually true, the gain is: -10

DESIGN MODEL

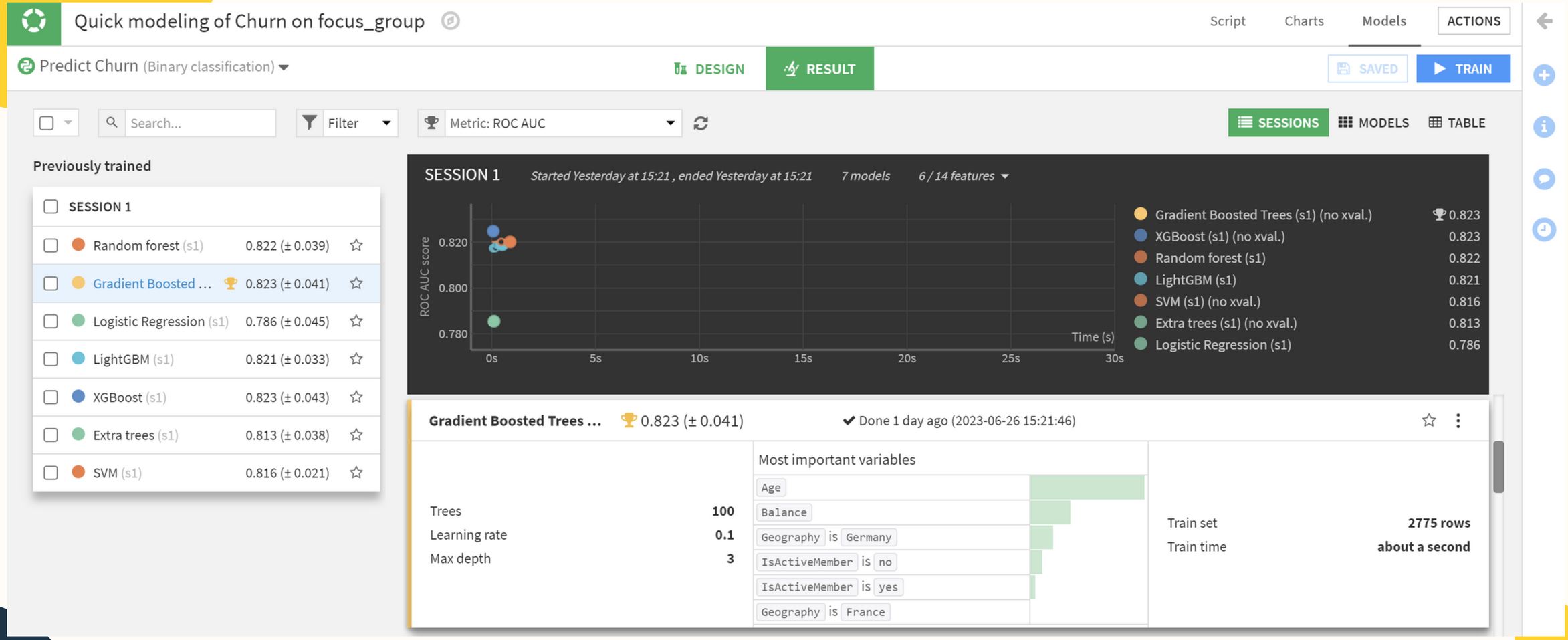
FEATURE HANDLING



ALGORITHMS SELECTION



TRAINING RESULT



TRAINING RESULT

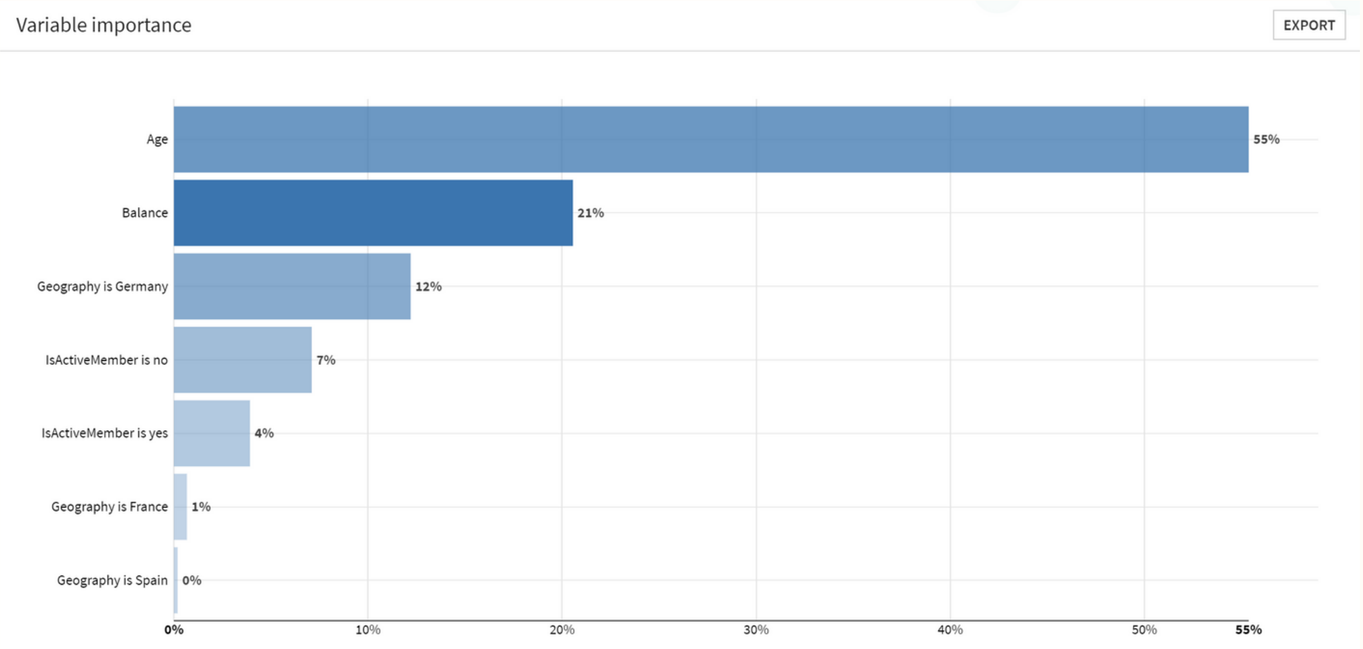
SUMMARY

ROC AUC: 0.823

Model

Model ID	**A-6004CHURNPREDICTION-wuOVSOh0-ew8AX0Ww-s1-pp5-m1**
Model type	**Two-class classification**
Target	**Churn**
Classes	**no** **yes**
Backend	**Python (in memory)**
Algorithm	**Gbt classification**
Trained on	**2023/06/26 15:21**
Columns	**14**
Data set rows	**2775**
Number of folds	**5**
Calibration method	**No calibration**
Code Env	**DSS builtin env**
Python version	**3.7.13**

VARIABLE IMPORTANT



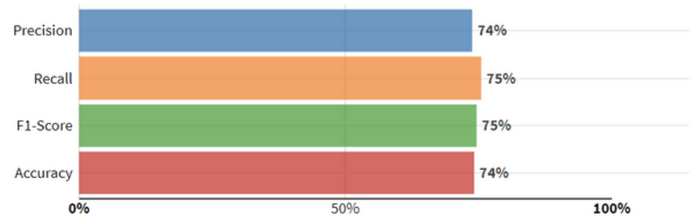
TRAINING RESULT

CONFUSION & COST METRIX

Confusion matrix

Display: Record count

	Predicted yes	Predicted no	Total
Actually yes	206	69	275
Actually no	76	204	280
Total	282	273	555

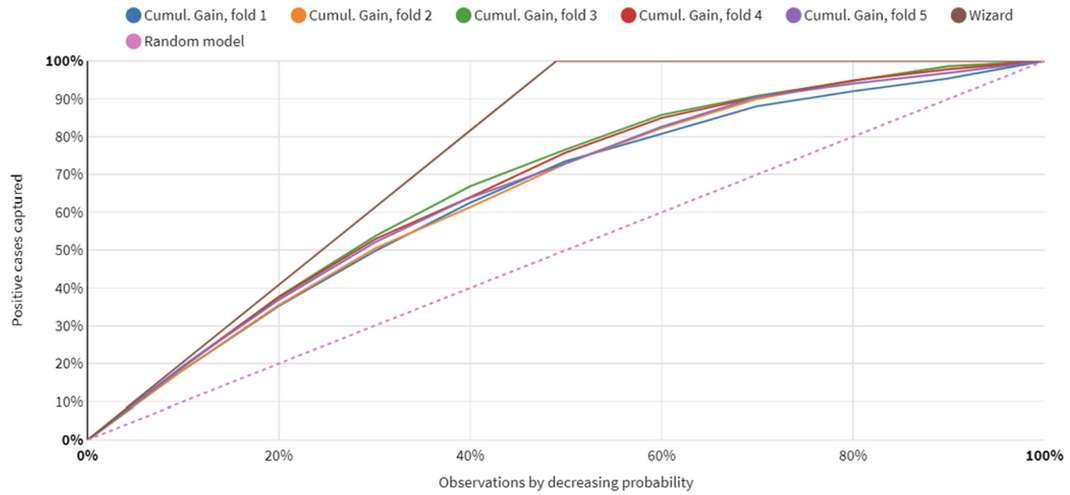


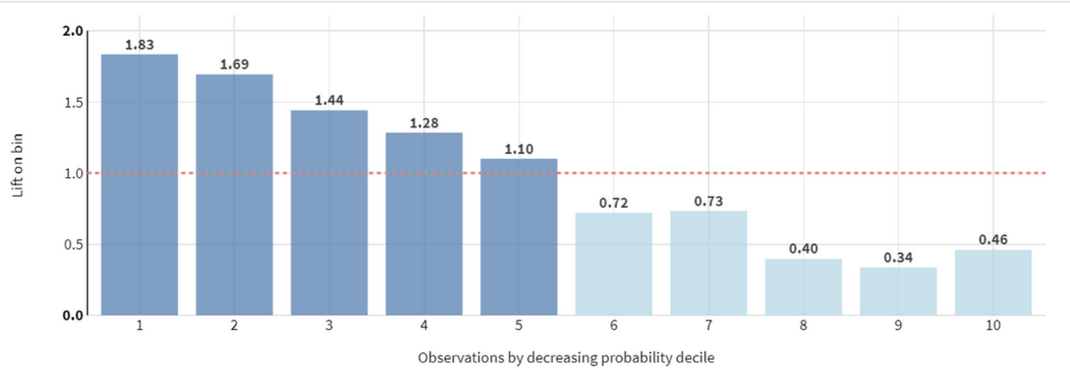
Cost matrix

If model predicts yes	and value is yes	the gain is	-3	×	206	=	-618.00
	but value is no	the gain is	-3	×	76	=	-228.00
Model predicts no	and value is no	the gain is	0	×	204	=	0.00
	but value is yes	the gain is	-10	×	69	=	-690.00
	Average gain per record		**-2.77**	×	555	=	-1,536.00

LIFT CHART

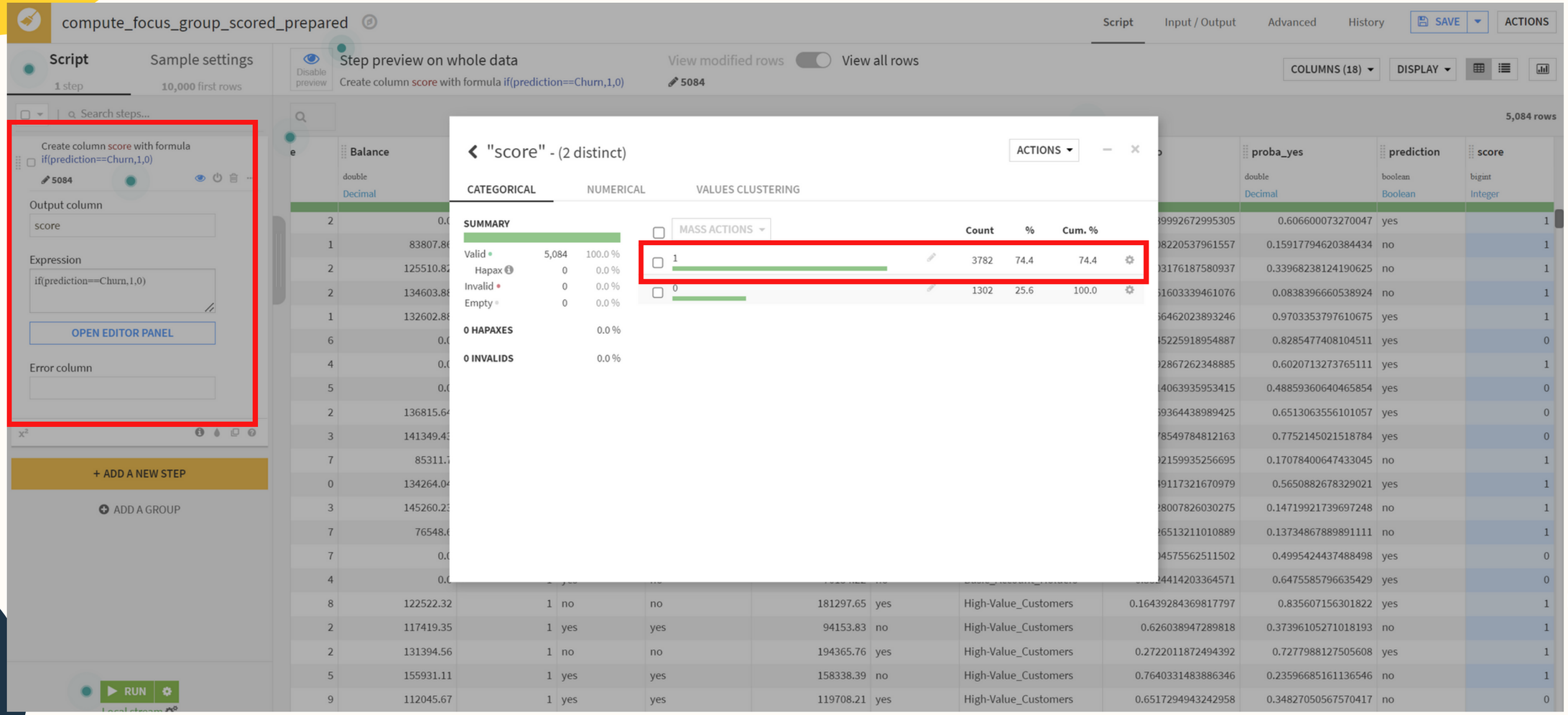
Lift charts



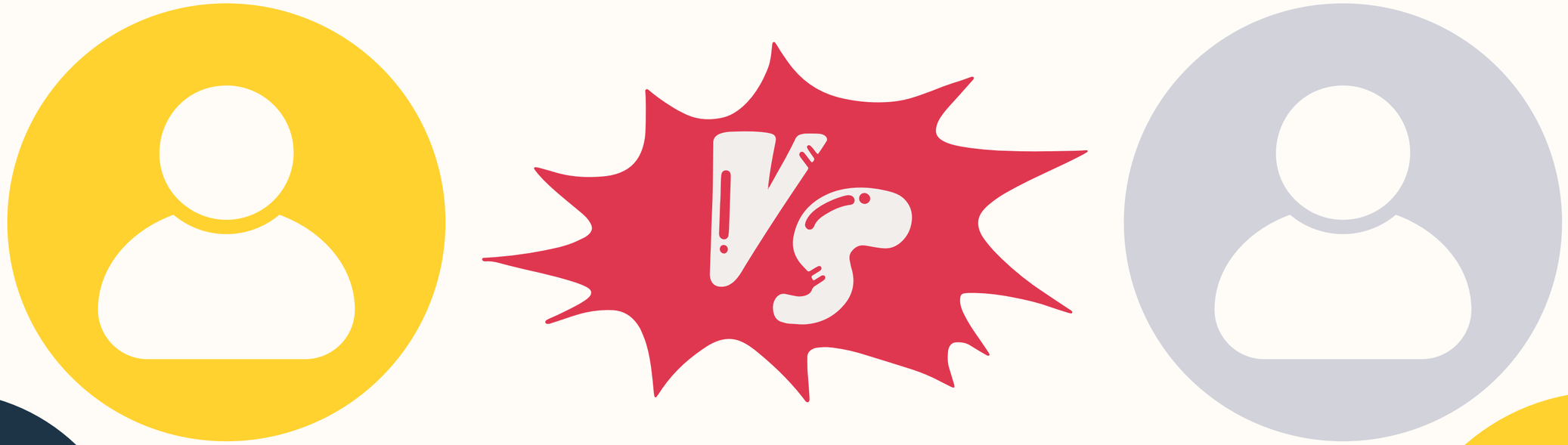


TRAINING RESULT

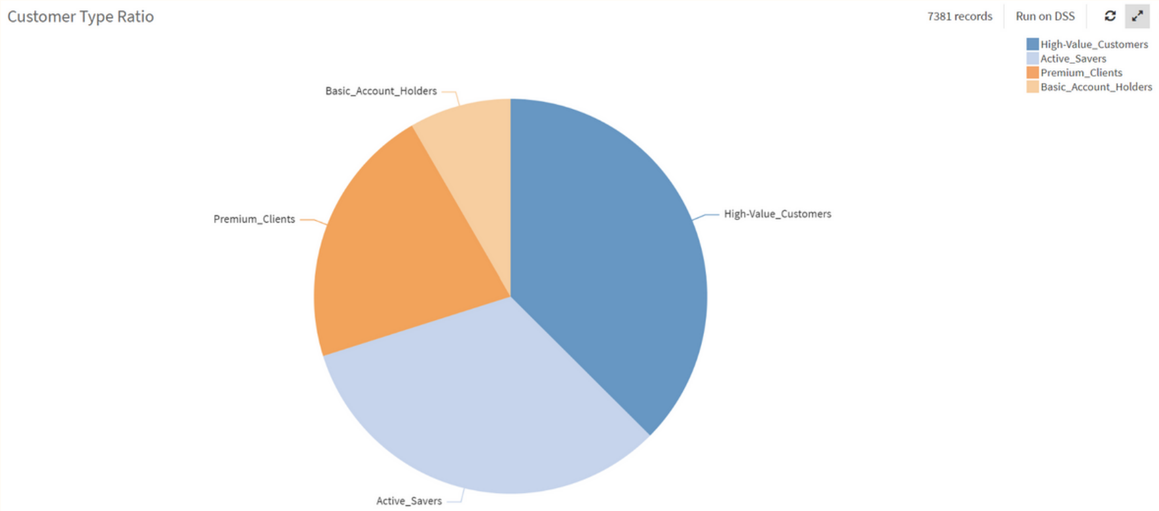
PREDICTION IN TRINING DATASET, CORRECT = 74.4%

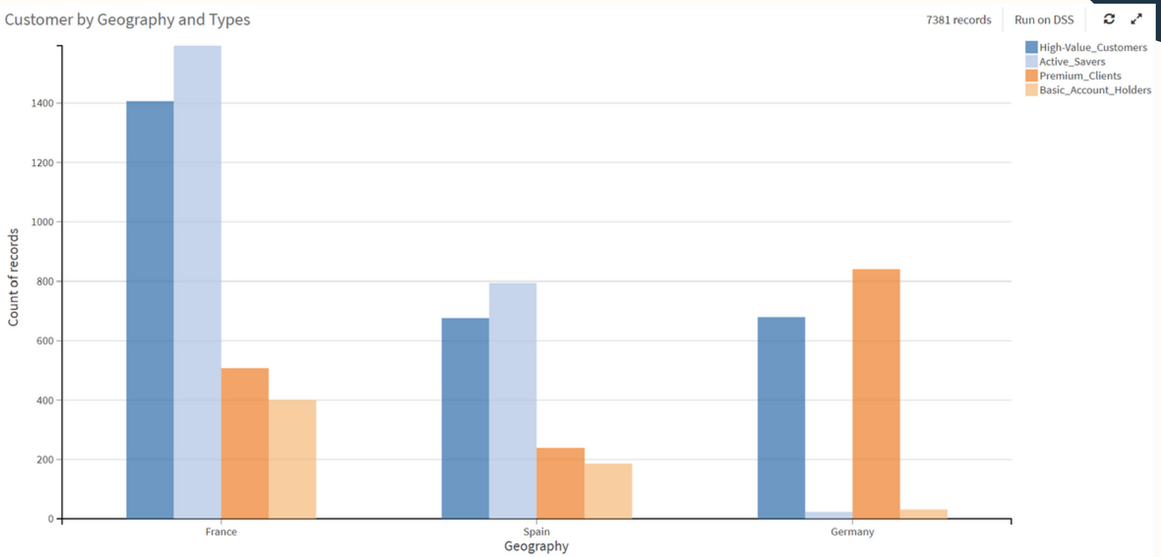


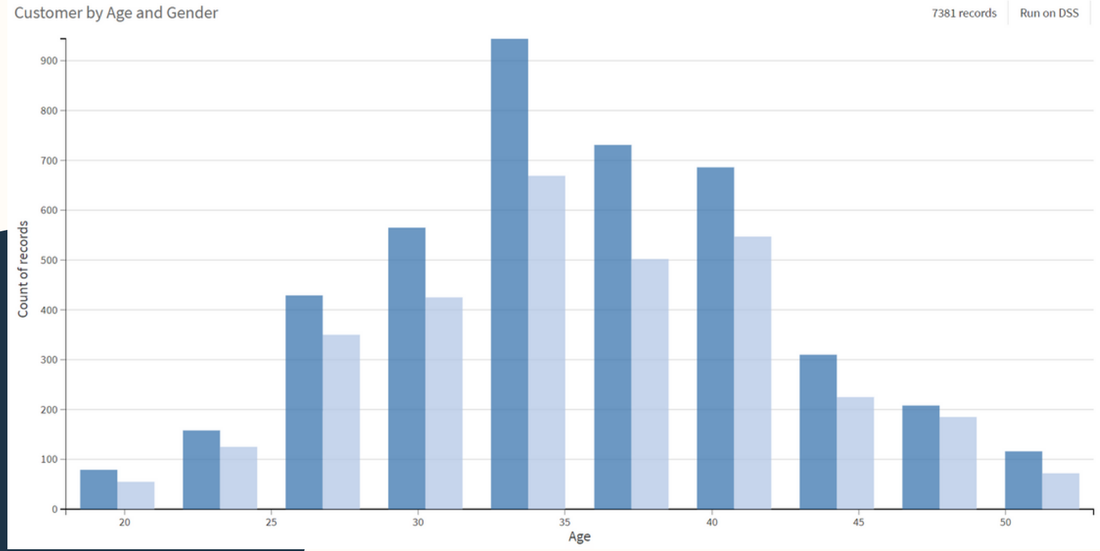
PART 3: CUSTOMER & CHURNER PROFILE

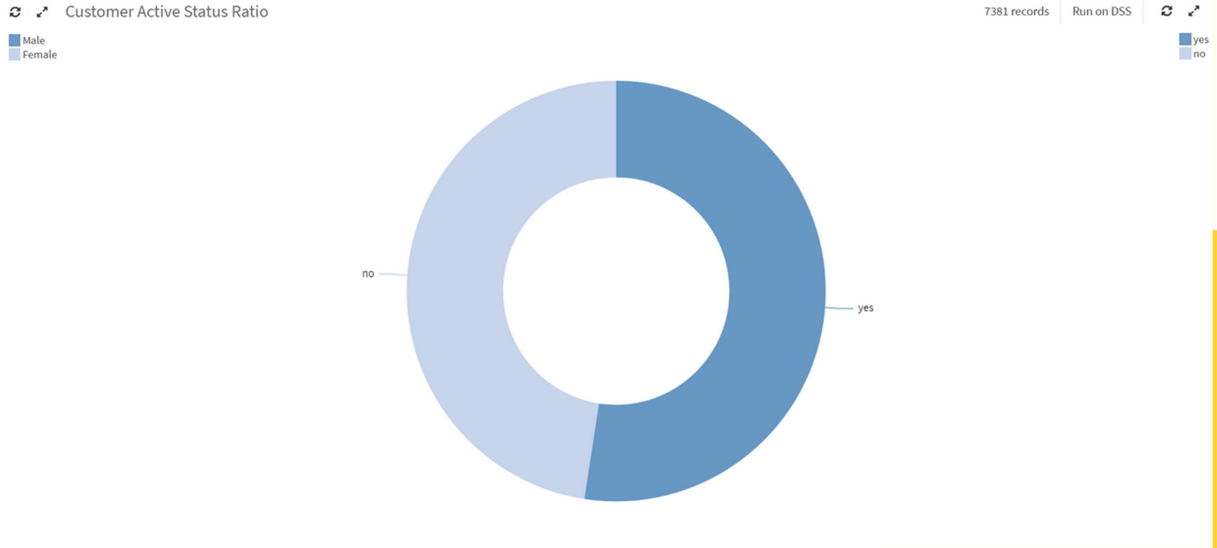


CUSTOMER PROFILE

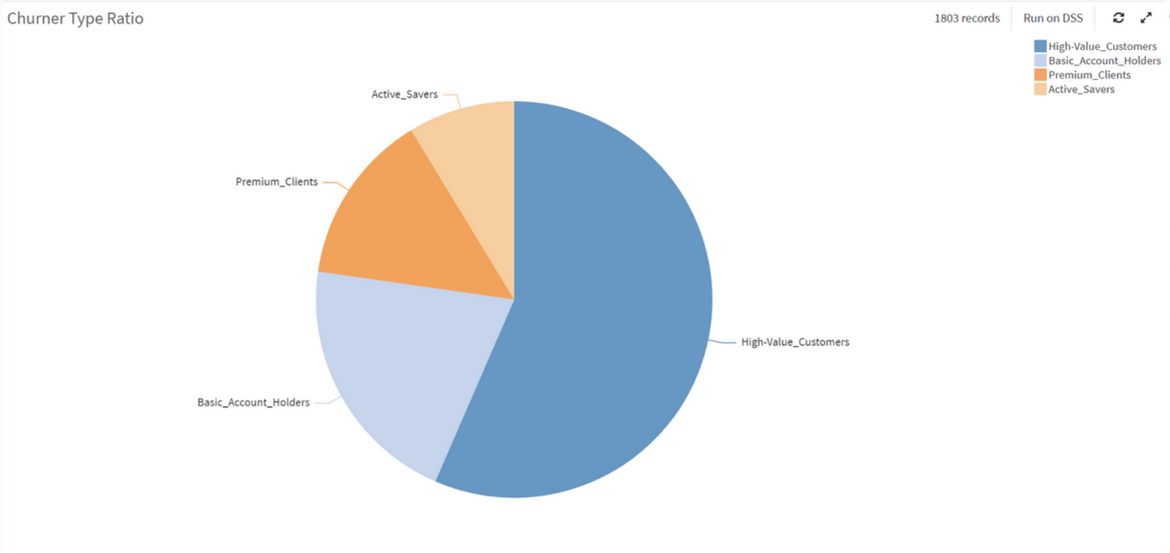


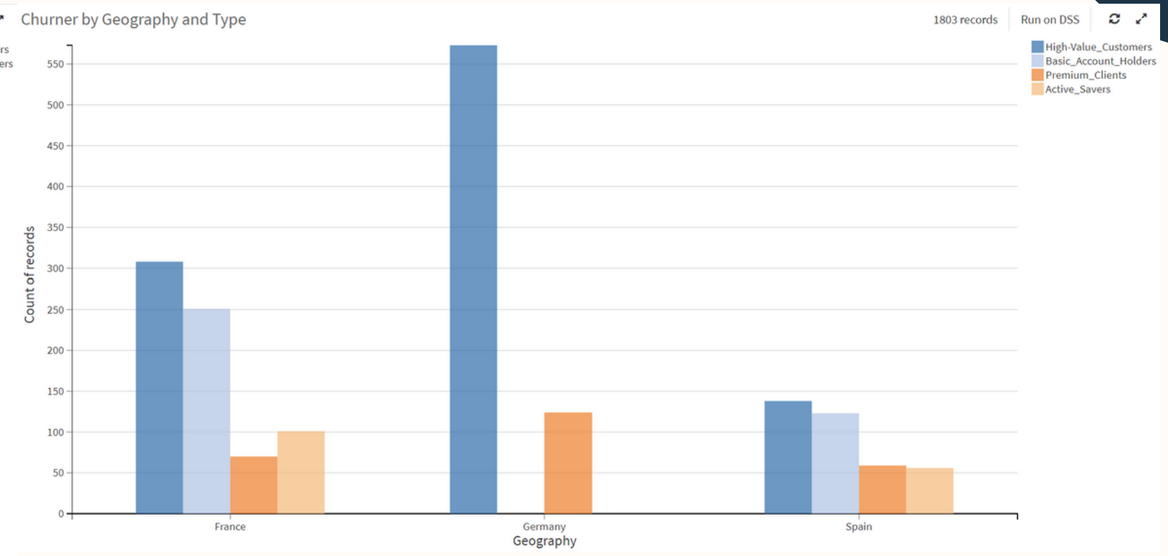


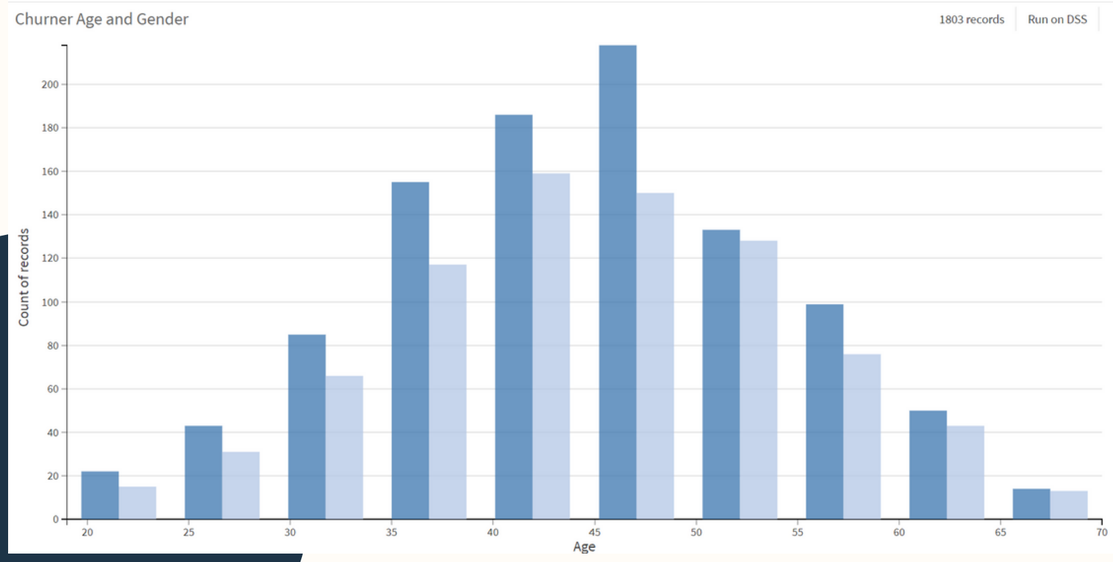


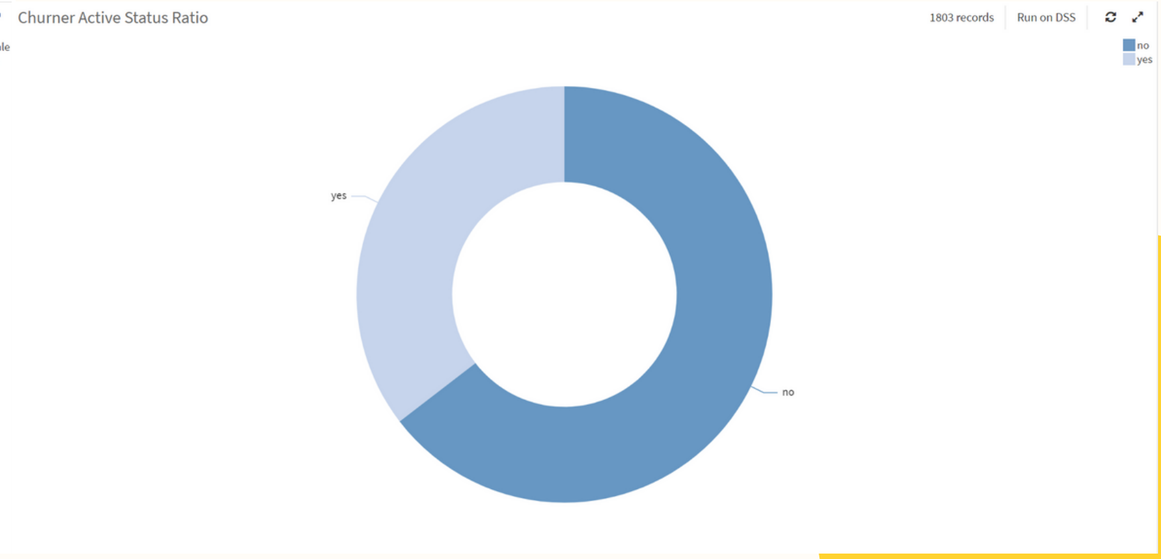


CHURNER PROFILE









PART 5: BUSINESS IMPACT

Impact 1

Based on customer group, Active Savers has low churining ratio which is 6% only. Business may upslling or crosss-selling to Basic Account Holders to increase no. of product and turn them into Active Savers group.

Impact 2

Retention program to reduce churning rate in focus group which are High-Value Customers and Basic Account Holders

Impact 3

Cost saving and reduce opportunity loss from confusion & cost matrix (explaination next slide)

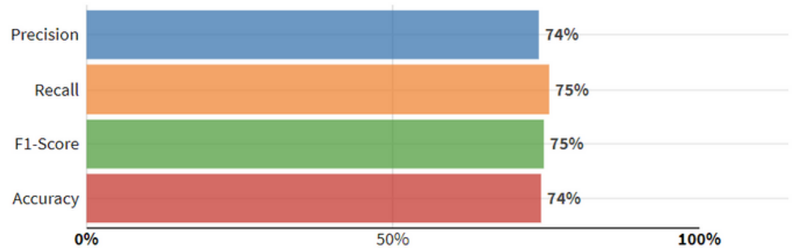
IMPACT 3 EXPLAINATION

CONFUSION & COST METRIX

Confusion matrix

Display: Record count

	Predicted yes	Predicted no	Total
Actually yes	206	69	275
Actually no	76	204	280
Total	282	273	555



Cost matrix

If model predicts yes	and value is yes	the gain is	-3	× 206	=	-618.00
	but value is no	the gain is	-3	× 76	=	-228.00
Model predicts no	and value is no	the gain is	0	× 204	=	0.00
	but value is yes	the gain is	-10	× 69	=	-690.00
	Average gain per record		**-2.77**	× 555	=	-1,536.00

Assumption

Uising ML to predict churn

- Cost to maintain customers which model predict churn = yes (Cost = 3)
- Opportunity loss from losing customers = 10

in this case of using ML to predict churn all cost & opportunity loss occur 1,536.00

Not using ML to predict churn and belive all there are loyalty customers, opportunity loss occur = 275(churned customers) * 10(opportunity loss/churn) = 2,750.00

Conclusion if use ML to predict churn will reduce opportunity loss around 2,750.00 - 1536.00 = 1,214.00


Thank
You