ibi Golden Summit

How WebFOCUS DSML Capabilities simplify your business processes

Ayden Ellsmere
Senior Solutions Engineer, ibi

Marcus Horbach
Principal Data Scientist, ibi

Confidentiality & Disclaimer

The information in this document is confidential information of Cloud Software Group, Inc. and/or its affiliates. Use, duplication, transmission, or republication for any purpose without the prior written consent of Cloud Software Group, Inc. is expressly prohibited.

This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release and availability dates for Cloud Software Group, Inc. products and services. This document is provided for informational purposes only and its contents are subject to change without notice. Cloud Software Group, Inc. makes no warranties, express or implied, in or relating to this document or any information in it, including, without limitation, that this document, or any information in it, is error-free or meets any conditions of merchantability or fitness for a particular purpose.

The material provided is for informational purposes only, and should not be relied on in making a purchasing decision. The information is not a commitment, promise or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any features or functionality described for our products remains at our sole discretion.

During the course of this presentation, Cloud Software Group, Inc. or its representatives may make forward-looking statements regarding future events Cloud Software Group, Inc.'s future results or our future financial performance. These statements are based on management's current expectations. Although we believe that the expectations reflected in the forward-looking statements contained in this presentation are reasonable, these expectations or any such forward-looking statements could prove to be incorrect and actual results or financial performance could differ materially from those stated herein. Cloud Software Group, Inc. does not undertake to update any forward-looking statement that may be made from time to time or on its behalf.

Agenda

- **Machine Learning Functions**

  Predictive analytics: train and use machine learning models
  Inside the core of the reporting engine: TABLE / COMPUTE
  Model-predictions generated inside your reports
  Explainability: clarify predictions + prescriptive analytics

- **Instant Insights**

  Shows insights (visualizations) into the data with just one click on a button

- **Natural Language Query**

  Allows you to query the data by asking English-language questions

- **Installation**

  Access DSML functionality: install binaries and inform profiles

Data Science and Machine Learning (DSML-package)

in WebFOCUS

Machine Learning Functions

Train and use Machine Learning Models in WebFOCUS

Classification ●
Regression ●
Anomaly Detection ○
Clustering ○
Time-Series Forecasting ●

50 YEARS OF INNOVATION

Information Builders

Time-Series Forecasting

Example:
. Datetime column with transaction timestamps
. Column with ‘units sold’ of some product per transaction

Question:
. Can you forecast to number of units sold per month in the future?
. Univariate time-series forecasting

We have to be a more precise:
. Time-series: observations made at <u>regular time intervals</u>, but the transaction times usually are not at regular intervals
. Forecasting methods need time-series . . .

The way to solve this:

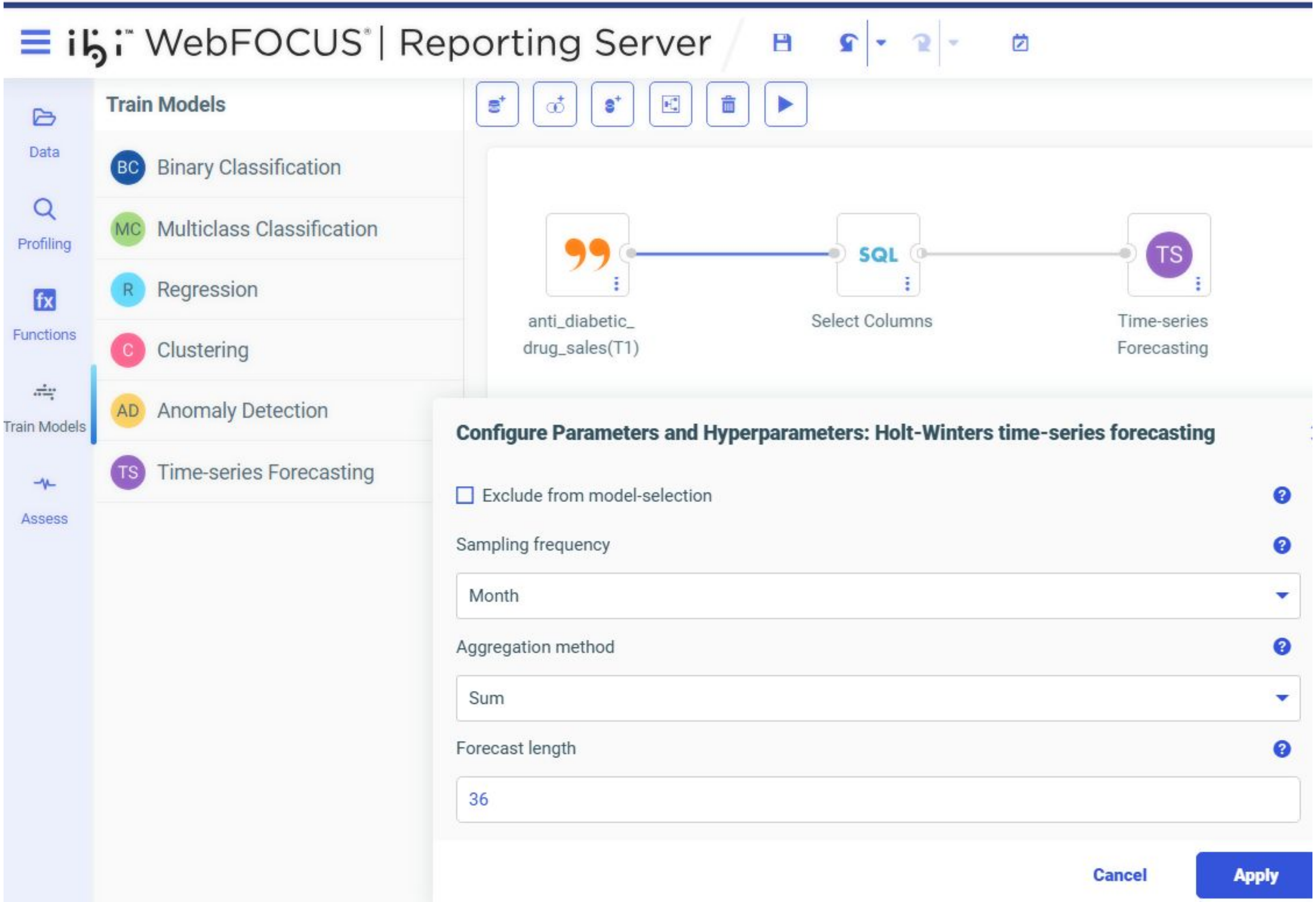
1. **Create a time-series**: (SUM) (MONTH)
. Convert data to regular time-intervals with **aggregation + resampling**: 'monthly total number of units sold'.

2. **Investigate the time-series for patterns**:
. Holt-Winters, Facebook-Prophet, SARIMA, Neural Network

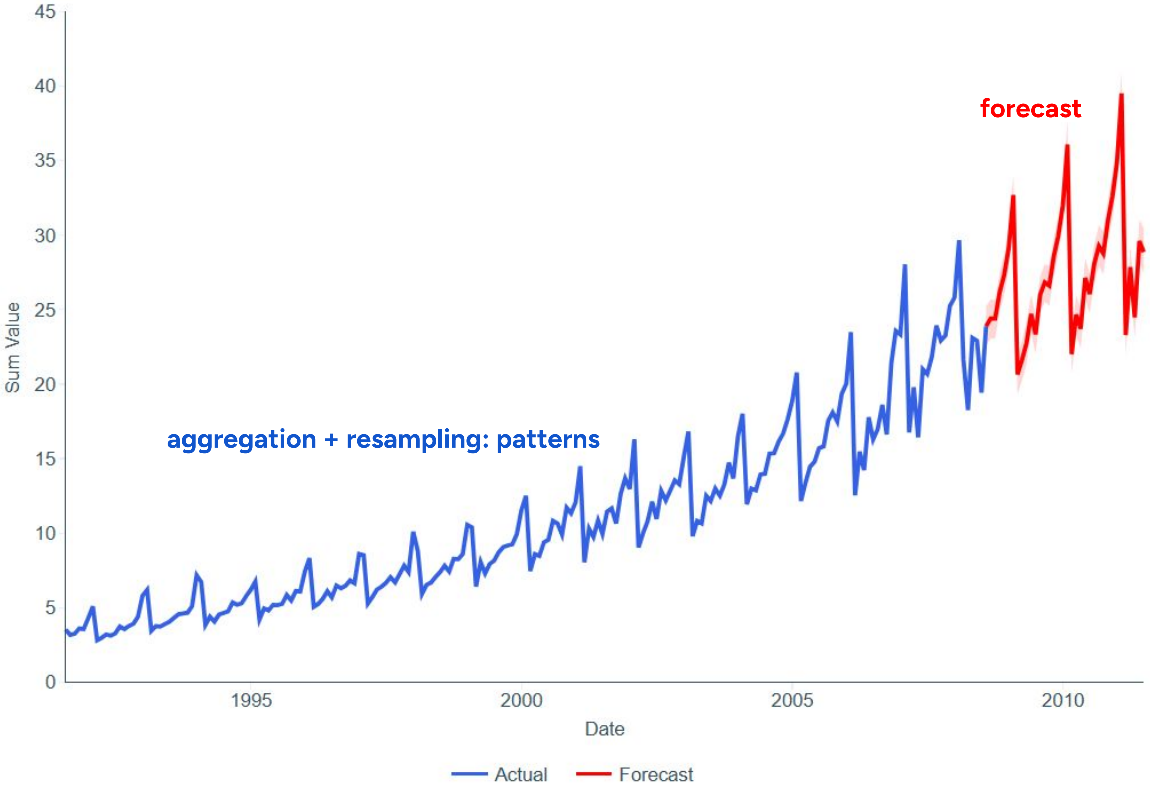
3. **Exploit the detected patterns to create forecasts**:
. Under the assumption that the detected patterns remain valid, forecasts can be made.

In WebFOCUS ML Functions:
FORECAST_HW, FORECAST_PROPHET, FORECAST_SARIMA, FORECAST_DNN

Information Builders



Information Builders


forecast
aggregation + resampling: patterns
Sum Value
45
40
35
30
25
20
15
10
5
0
1995
2000
2005
2010
Date
Actual
Forecast

Supervised Machine Learning for tabular data

Binary Classification, Multiclass Classification, Regression

. Understand the values in one chosen column (target y)

in terms of the values in some or all of other columns (predictors X)

. That is: find a function F for which $y \cong F(X)$

. F is called a model, it can be applied to new data X_new for making 'predictions': y_pred = F(X_new)

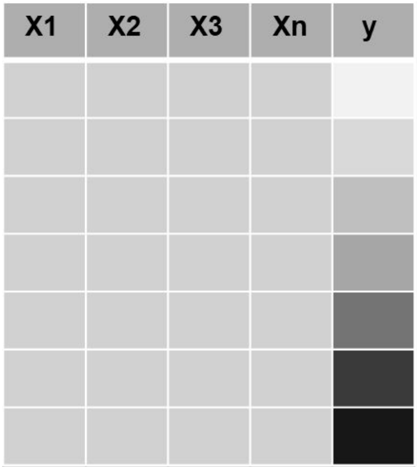
. F can be a 'calculus' type of function, or a decision tree, or a collection of decision trees, or a neural network

. Various algorithms to find F (Logistic Regression, Random Forest, XGBoost, etc)

. Finding F is called 'training the model', and (X,y) is called the training-data

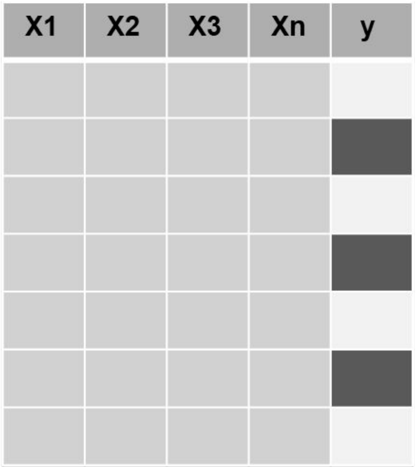
Regression

continuous target values



Binary Classification

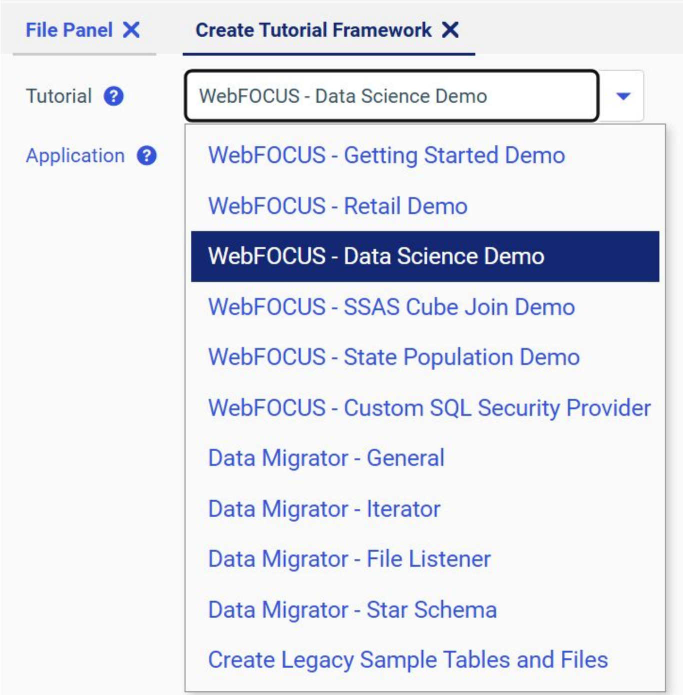
binary target values



Multiclass Classification

discrete set of (unordered) target labels

Explore **Data Science Demo Tutorial** for various examples of regression and classification models.



A Binary Classification Example: Employee Attrition Prediction

. training
. inference
. explainability

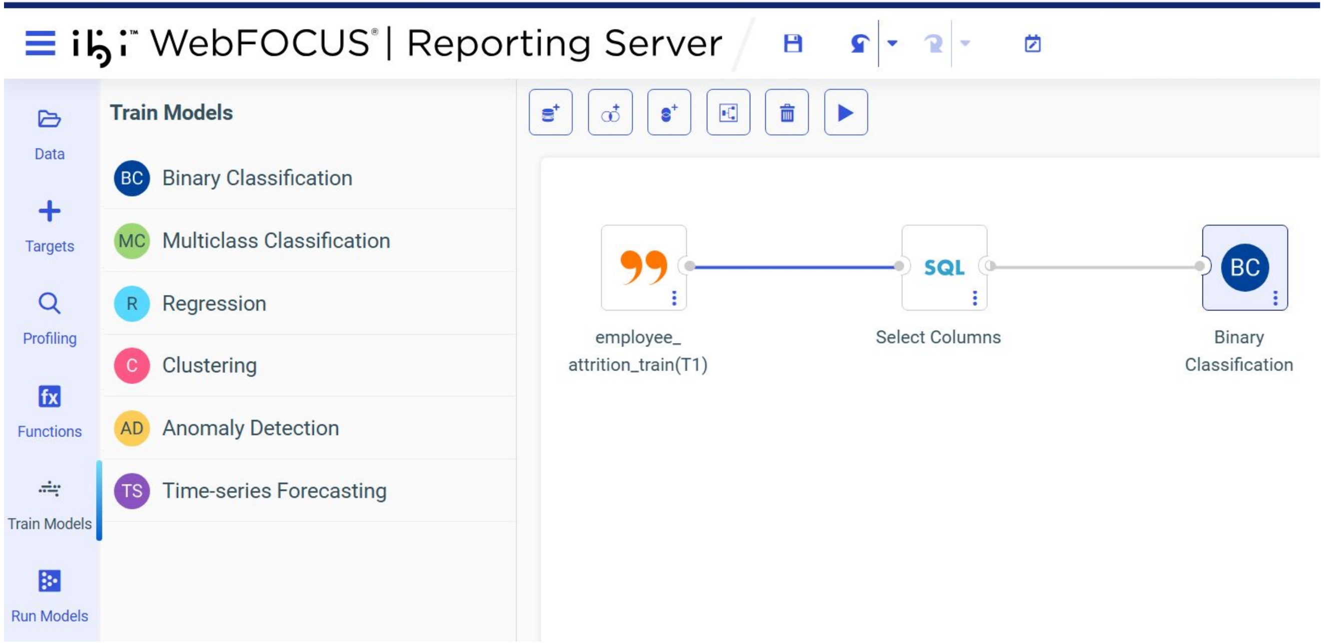
Employee Attrition Data: demographic + job related + attrition flag

Employee ID	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	Overtime	Distance from Home	Education Level	Marital Status
8410	31	Male	19	Education	5390	Excellent	Medium	Average	2	No	22	Associate Degree	Married
64756	59	Female	4	Media	5534	Poor	High	Low	3	No	21	Master Degree	Divorced
30257	24	Female	10	Healthcare	8159	Good	High	Low	0	No	11	Bachelor Degree	Married
65791	36	Female	7	Education	3989	Good	High	High	1	No	27	High School	Single
65026	56	Male	41	Education	4821	Fair	Very High	Average	0	Yes	71	High School	Divorced
24368	38	Female	3	Technology	9977	Fair	High	Below Average	3	No	37	Bachelor Degree	Married

Number of Dependents	Job Level	Company Size	Company Tenure	Remote Work	Leadership Opportunities	Innovation Opportunities	Company Reputation	Employee Recognition	Attrition
0	Mid	Medium	89	No	No	No	Excellent	Medium	Stayed
3	Mid	Medium	21	No	No	No	Fair	Low	Stayed
3	Mid	Medium	74	No	No	No	Poor	Low	Stayed
2	Mid	Small	50	Yes	No	No	Good	Medium	Stayed
0	Senior	Medium	68	No	No	No	Fair	Medium	Stayed
0	Mid	Medium	47	No	No	Yes	Fair	High	Left

Employee Attrition

1. Create a Data Flow for training a Binary Classification model

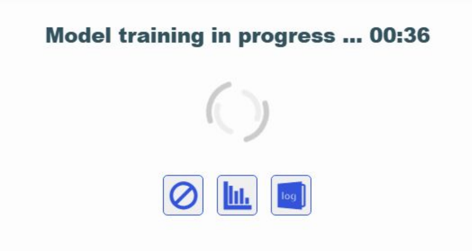


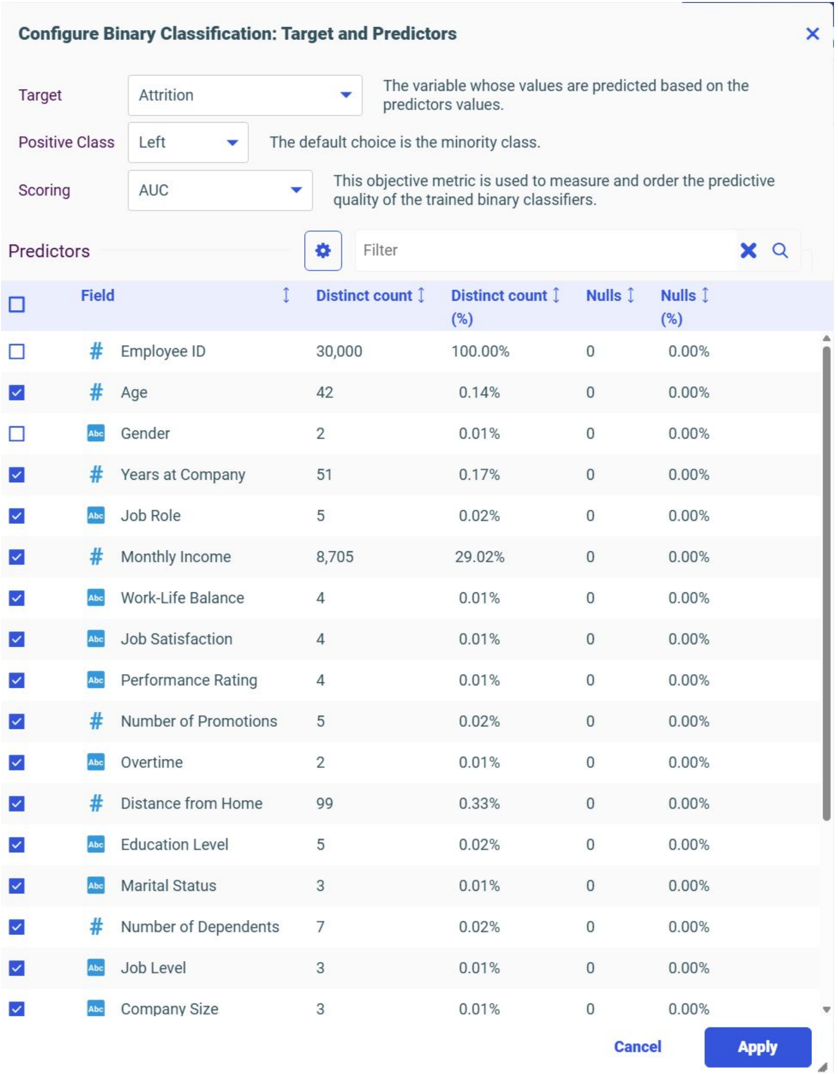
Employee Attrition

2. Specify the problem

. Choose the target
. Choose the predictors

. Click 'Train'





Configure Binary Classification: Target and Predictors

Target: Attrition — The variable whose values are predicted based on the predictors values.

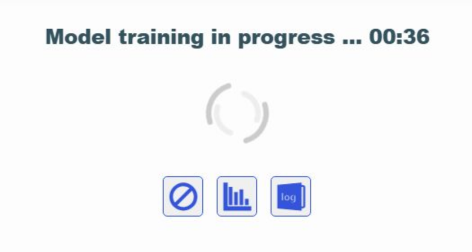
Positive Class: Left — The default choice is the minority class.

Scoring: AUC — This objective metric is used to measure and order the predictive quality of the trained binary classifiers.

Predictors — Filter

	Field	Distinct count	Distinct count (%)	Nulls	Nulls (%)
☐	Employee ID	30,000	100.00%	0	0.00%
☑	Age	42	0.14%	0	0.00%
☐	Gender	2	0.01%	0	0.00%
☑	Years at Company	51	0.17%	0	0.00%
☑	Job Role	5	0.02%	0	0.00%
☑	Monthly Income	8,705	29.02%	0	0.00%
☑	Work-Life Balance	4	0.01%	0	0.00%
☑	Job Satisfaction	4	0.01%	0	0.00%
☑	Performance Rating	4	0.01%	0	0.00%
☑	Number of Promotions	5	0.02%	0	0.00%
☑	Overtime	2	0.01%	0	0.00%
☑	Distance from Home	99	0.33%	0	0.00%
☑	Education Level	5	0.02%	0	0.00%
☑	Marital Status	3	0.01%	0	0.00%
☑	Number of Dependents	7	0.02%	0	0.00%
☑	Job Level	3	0.01%	0	0.00%
☑	Company Size	3	0.01%	0	0.00%

Cancel Apply



What is happening during the training ?

Find $y \cong F(X)$ from the training data (X, y)

> Our models are complete pipelines

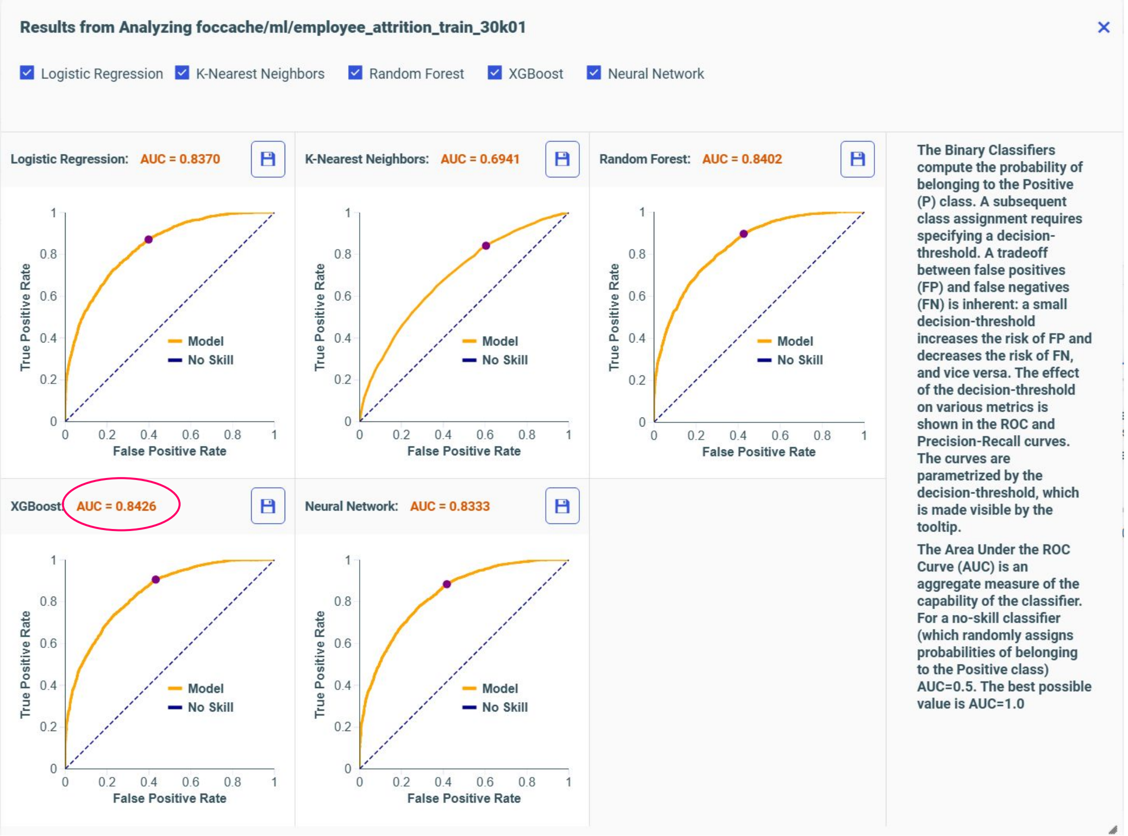
. containing preprocessing + machine learning steps
. . preprocessing: encoding of non-numeric columns, incl. dates
. . preprocessing: imputation of missing values
. . machine learning: models of 5 different types (algorithms) are trained

> Statistical responsibility

. detect relevant patterns but not accidental patterns
. balancing act controlled using regularization techniques involving hyperparameters
. automated hyperparameter-optimization through various strategies (grid-search)
. model-evaluation

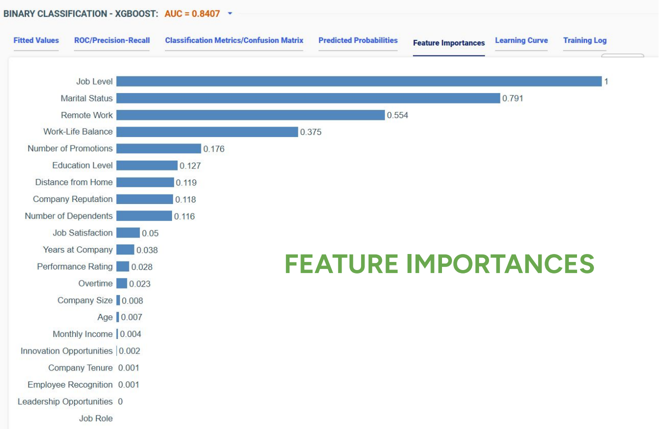
Employee Attrition

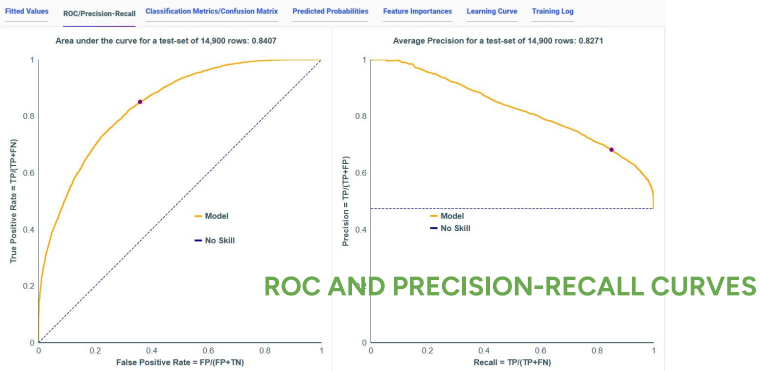
3. Evaluate the trained models

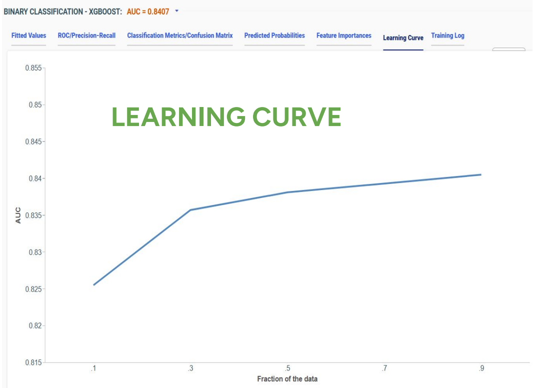


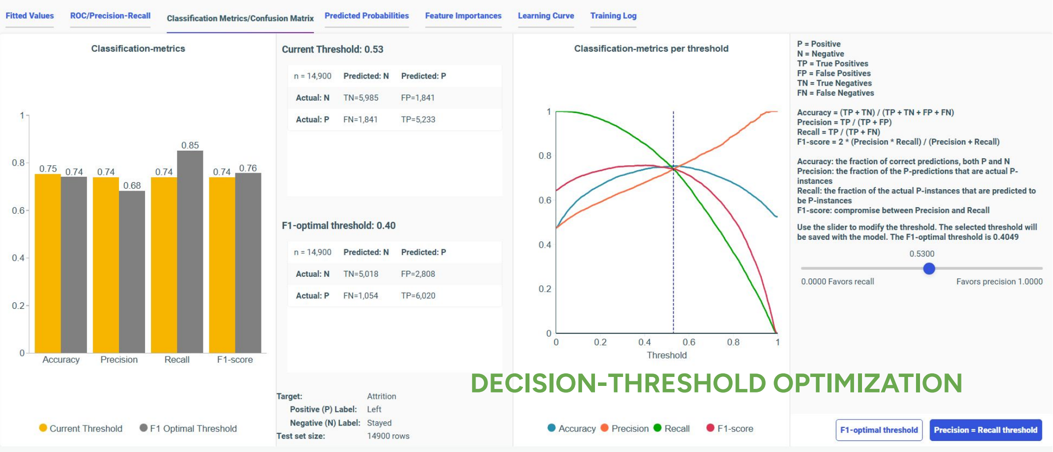
Employee Attrition

3*. Evaluate the trained models on specialist level



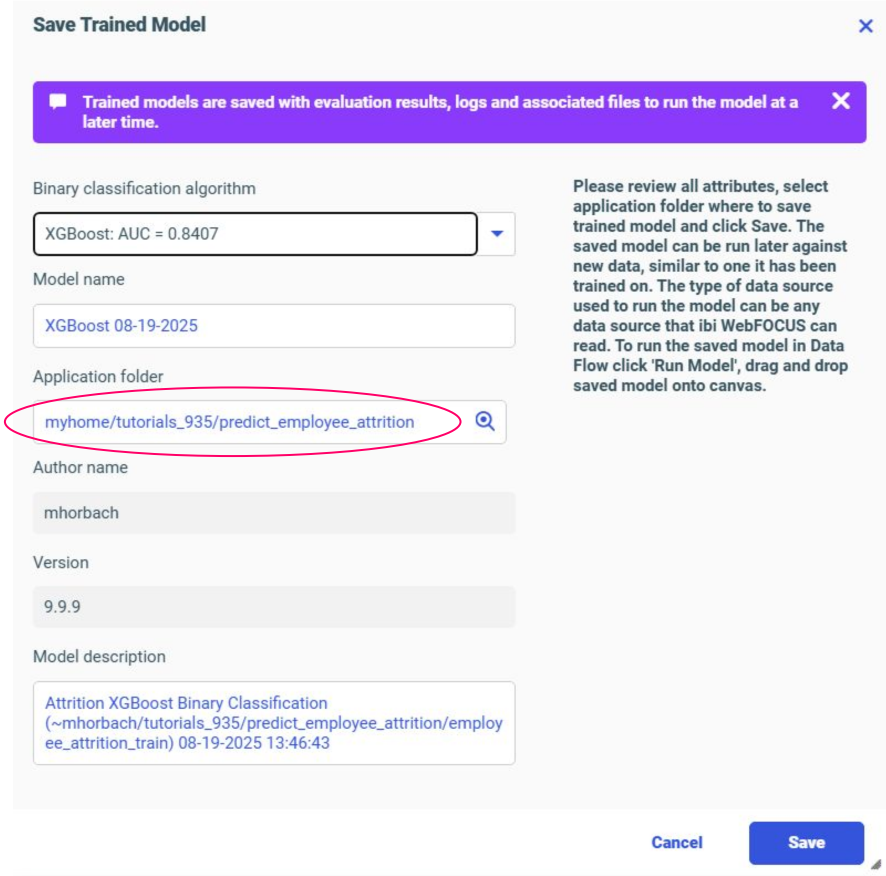






Employee Attrition

4. Save the best model



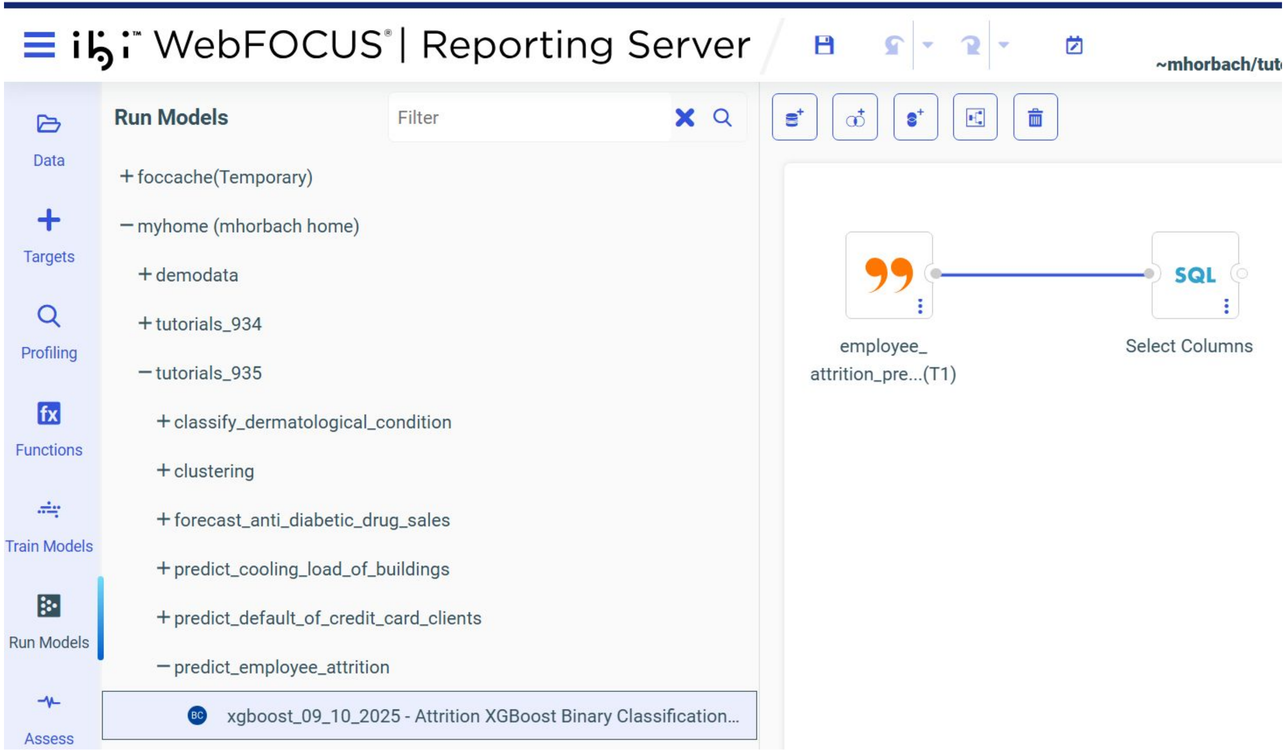
Employee Attrition Inference, i.e. use the saved model to make predictions for NEW data

Employee ID	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	Overtime	Distance from Home	Education Level	Marital Status
48210	26	Male	6	Finance	7317	Poor	High	Average	0	Yes	51	Bachelor Degree	Single
8438	28	Male	3	Media	6535	Fair	Very High	Average	2	Yes	23	Associate Degree	Married
53584	33	Male	5	Technology	8904	Fair	Very High	Below Average	2	No	85	Associate Degree	Single
44128	20	Male	4	Technology	11249	Excellent	Very High	Low	0	No	57	Associate Degree	Single
30115	59	Male	34	Technology	8268	Excellent	Low	Average	1	Yes	71	Master Degree	Married

Marital Status	Number of Dependents	Job Level	Company Size	Company Tenure	Remote Work	Leadership Opportunities	Innovation Opportunities	Company Reputation	Employee Recognition
Single	1	Mid	Large	23	No	No	No	Good	High
Married	3	Mid	Medium	76	Yes	No	No	Fair	Medium
Single	2	Mid	Medium	6	No	No	Yes	Good	Medium
Single	3	Entry	Small	55	Yes	No	No	Good	Low
Married	0	Senior	Medium	91	No	No	No	Poor	Medium

Employee Attrition

5. Create a Data Flow for running a saved Binary Classification model



Employee Attrition 6. Inference

Attrition Predicted	Attrition Predicted Probability	Employee ID	Age	Gender	Years at Company	Job Role	Monthly Income
Left	.9368	48210	26	Male	6	Finance	73
Left	.9304	53584	33	Male	5	Technology	89
Left	.7035	44128	20	Male	4	Technology	112
Stayed	.3048	8438	28	Male	3	Media	65
Stayed	.2276	30115	59	Male	34	Technology	82

Explainability at inference-time

Machine Learning models are increasingly being used for making decisions that directly affect humans

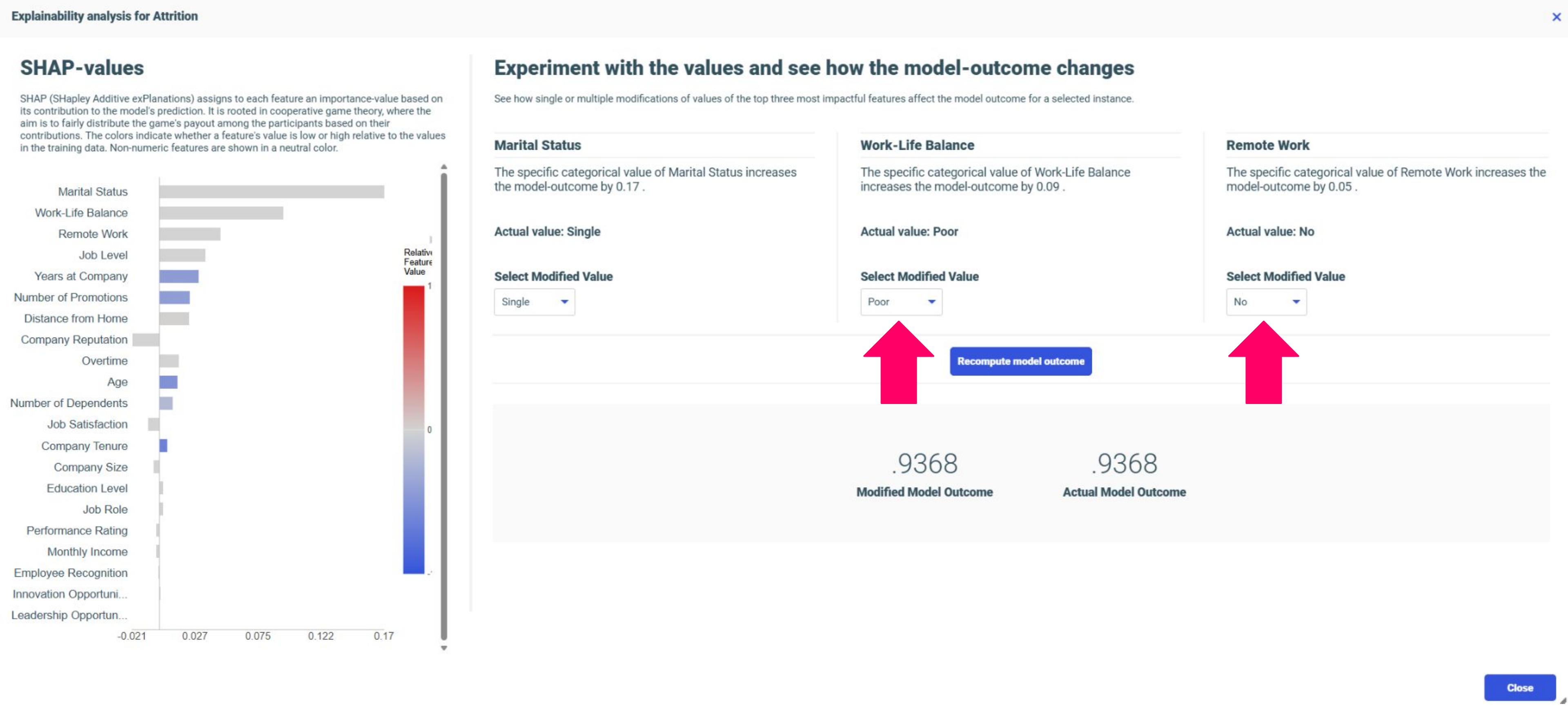
This raises questions of **trust and accountability**.

Provide supporting information about individual model outcomes:
Why does the model assess a loan-applicant's risk-of-default as high?

Explainability methods (SHAP, XWIN) provide such information in terms of the model's input-data

Offers 'intervention prescriptions':
. what to do to receive a more favorable model-outcome? (What can we do to have a loan application accepted)
. what to do to avoid some unwanted actual outcome? (What can we do to keep the employee)

Employee Attrition 7a. Explainability: the most influential factors

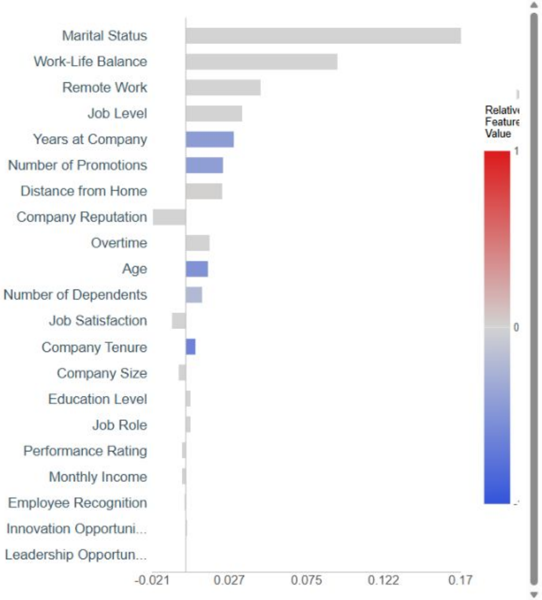


Employee Attrition 7b. Explainability: modify influential factors

Explainability analysis for Attrition

SHAP-values

SHAP (SHapley Additive exPlanations) assigns to each feature an importance-value based on its contribution to the model's prediction. It is rooted in cooperative game theory, where the aim is to fairly distribute the game's payout among the participants based on their contributions. The colors indicate whether a feature's value is low or high relative to the values in the training data. Non-numeric features are shown in a neutral color.



Experiment with the values and see how the model-outcome changes

See how single or multiple modifications of values of the top three most impactful features affect the model outcome for a selected instance.

Marital Status

The specific categorical value of Marital Status increases the model-outcome by 0.17 .

Actual value: Single

Select Modified Value

Single

Work-Life Balance

The specific categorical value of Work-Life Balance increases the model-outcome by 0.09 .

Actual value: Poor

Select Modified Value

Good

Remote Work

The specific categorical value of Remote Work increases the model-outcome by 0.05 .

Actual value: No

Select Modified Value

Yes

Recompute model outcome

.4599
Modified Model Outcome

.9368
Actual Model Outcome

Close

Ranking of FIFA-players (Regression Model)

Configure Regression: Target and Predictors

Target: Overall

The variable whose values are predicted based on the predictors values.

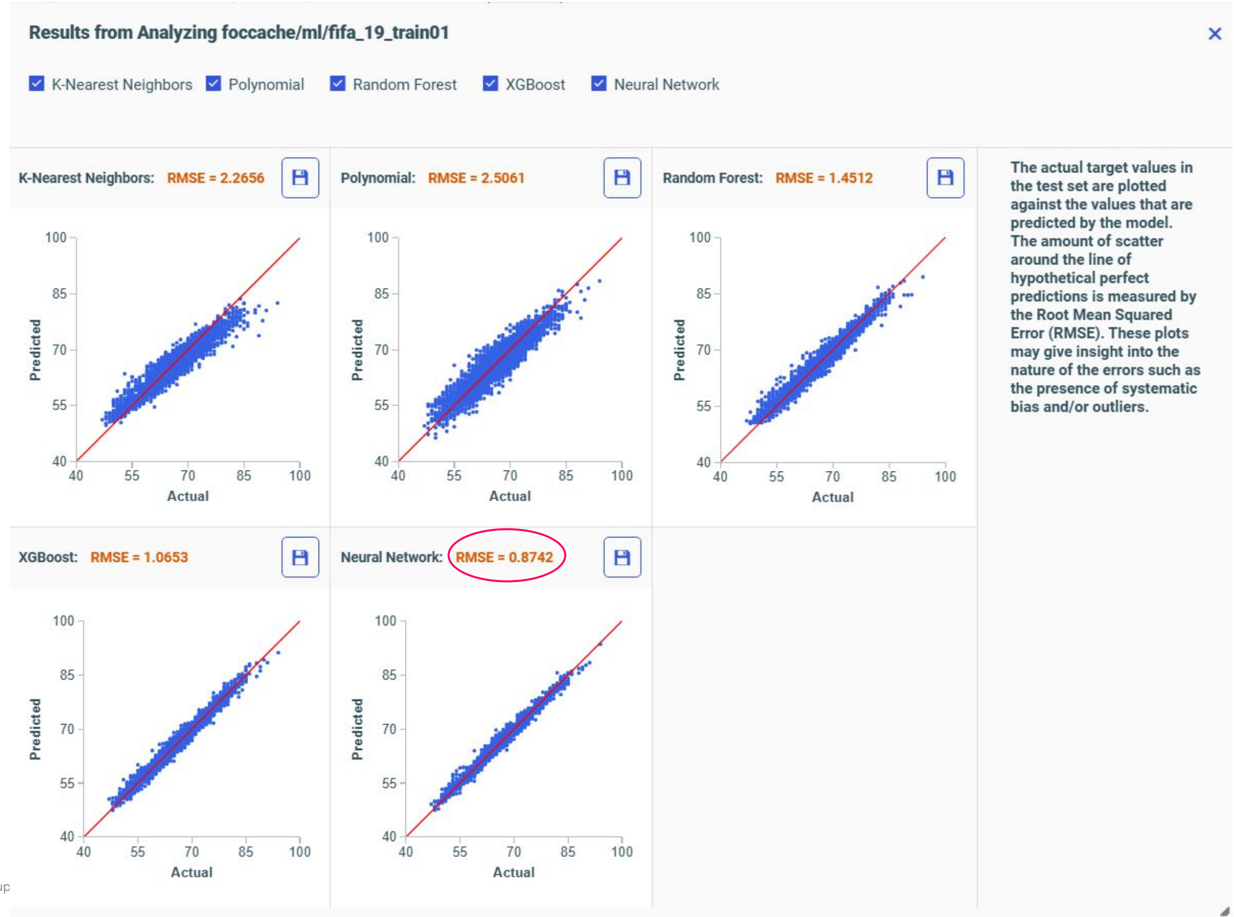
Predictors

Filter

	Field	Distinct count	Distinct count (%)	Nulls	Nulls (%)
☐	ID	17,936	98.81%	0	0.00%
☐	Name	16,894	93.07%	0	0.00%
☑	Age	29	0.16%	0	0.00%
☑	Preferred Foot	2	0.01%	0	0.00%
☑	Weak Foot	5	0.03%	0	0.00%
☑	Skill Moves	5	0.03%	0	0.00%
☑	Work Rate	9	0.05%	0	0.00%
☑	Body Type	8	0.04%	0	0.00%
☑	Position	27	0.15%	12	0.07%
☑	Jersey Number	100	0.55%	12	0.07%
☐	Joined	1,717	10.31%	1,504	8.29%
☑	Height	21	0.12%	0	0.00%
☑	Weight	57	0.31%	0	0.00%
☑	Crossing	89	0.49%	0	0.00%
☑	Finishing	93	0.51%	0	0.00%
☑	HeadingAccuracy	92	0.51%	0	0.00%
☑	ShortPassing	85	0.47%	0	0.00%
☑	Volleys	87	0.48%	0	0.00%
☑	Dribbling	95	0.52%	0	0.00%
☑	Curve	89	0.49%	0	0.00%

Cancel Apply

Ranking of FIFA-players (Regression)

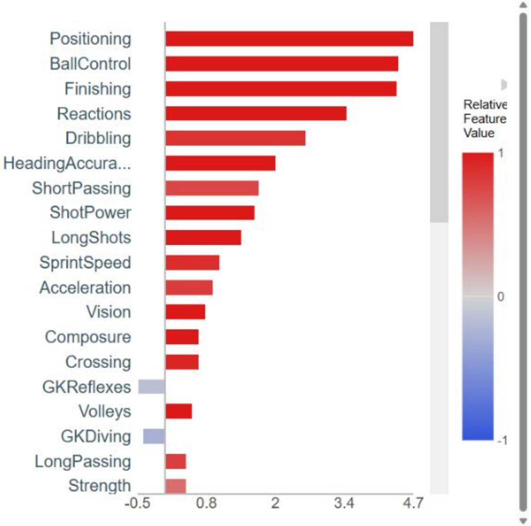


Neural network prediction for C. Ronaldo (a forward)

Explainability analysis for Overall

SHAP-values

SHAP (SHapley Additive exPlanations) assigns to each feature an importance-value based on its contribution to the model's prediction. It is rooted in cooperative game theory, where the aim is to fairly distribute the game's payout among the participants based on their contributions. The colors indicate whether a feature's value is low or high relative to the values in the training data. Non-numeric features are shown in a neutral color.



Experiment with the values and see how the model-outcome changes

See how single or multiple modifications of values of the top three most impactful features affect the model outcome for a selected instance.

Positioning

The very high value of Positioning increases the model-outcome by 4.7 .

Actual value: 95.0
Realistic Range: 2.0 to 94.0
Average value: 49.9

Enter Modified Value

95.0

BallControl

The very high value of BallControl increases the model-outcome by 4.42 .

Actual value: 94.0
Realistic Range: 5.0 to 96.0
Average value: 58.3

Enter Modified Value

94.0

Finishing

The very high value of Finishing increases the model-outcome by 4.38 .

Actual value: 94.0
Realistic Range: 2.0 to 95.0
Average value: 45.5

Enter Modified Value

94.0

Recompute model outcome

95
Modified Model Outcome

95
Actual Model Outcome

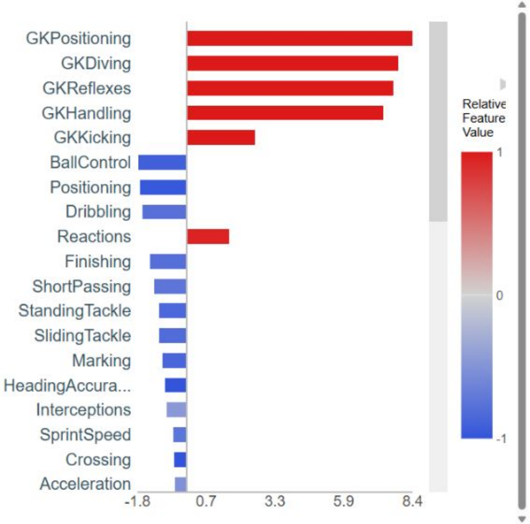
Close

Neural network prediction for G. Buffon (a goalkeeper)

Explainability analysis for Overall

SHAP-values

SHAP (SHapley Additive exPlanations) assigns to each feature an importance-value based on its contribution to the model's prediction. It is rooted in cooperative game theory, where the aim is to fairly distribute the game's payout among the participants based on their contributions. The colors indicate whether a feature's value is low or high relative to the values in the training data. Non-numeric features are shown in a neutral color.



Experiment with the values and see how the model-outcome changes

See how single or multiple modifications of values of the top three most impactful features affect the model outcome for a selected instance.

GKPositioning

The very high value of GKPositioning increases the model-outcome by 8.44 .

Actual value: 90.0
Realistic Range: 1.0 to 89.0
Average value: 16.3

Enter Modified Value

90.0

GKDiving

The very high value of GKDiving increases the model-outcome by 7.91 .

Actual value: 88.0
Realistic Range: 1.0 to 90.0
Average value: 16.6

Enter Modified Value

88.0

GKReflexes

The very high value of GKReflexes increases the model-outcome by 7.72 .

Actual value: 83.0
Realistic Range: 1.0 to 94.0
Average value: 16.7

Enter Modified Value

83.0

Recompute model outcome

86
Modified Model Outcome

86
Actual Model Outcome

Close

Instant Insights

. With a click on a button, you initiate the execution of a suite of statistical and machine learning algorithms, exploring the data.

. Running time from, roughly, 30 seconds to a few minutes.

. Interesting results are presented as visualizations.

. Individual visualizations can be moved to the canvas in Designer.

Instant Insights

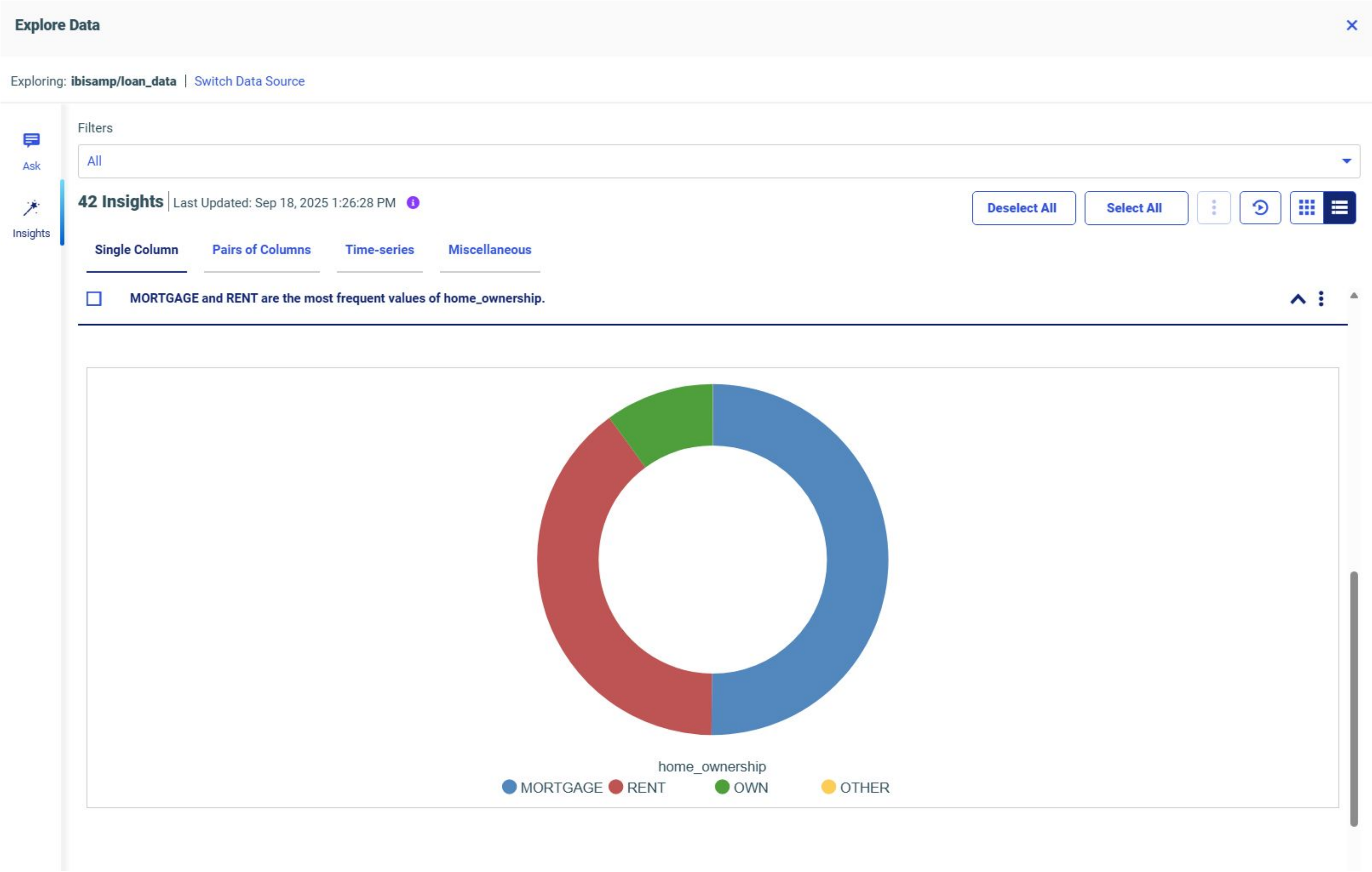
. With a click on a button, you initiate the execution of a suite of statistical and machine learning algorithms, exploring the data.

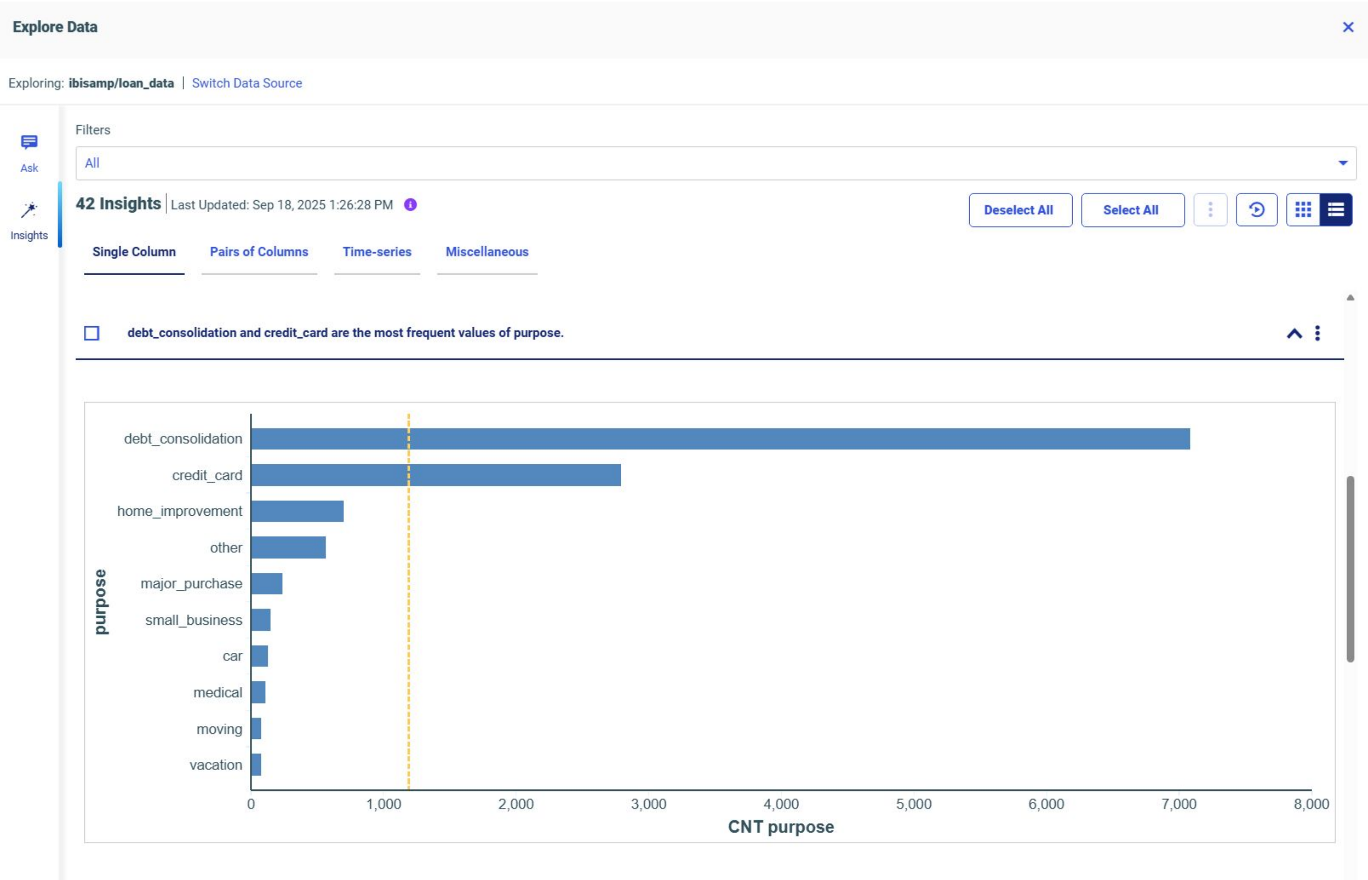
. Running time from, roughly, 30 seconds to a few minutes.

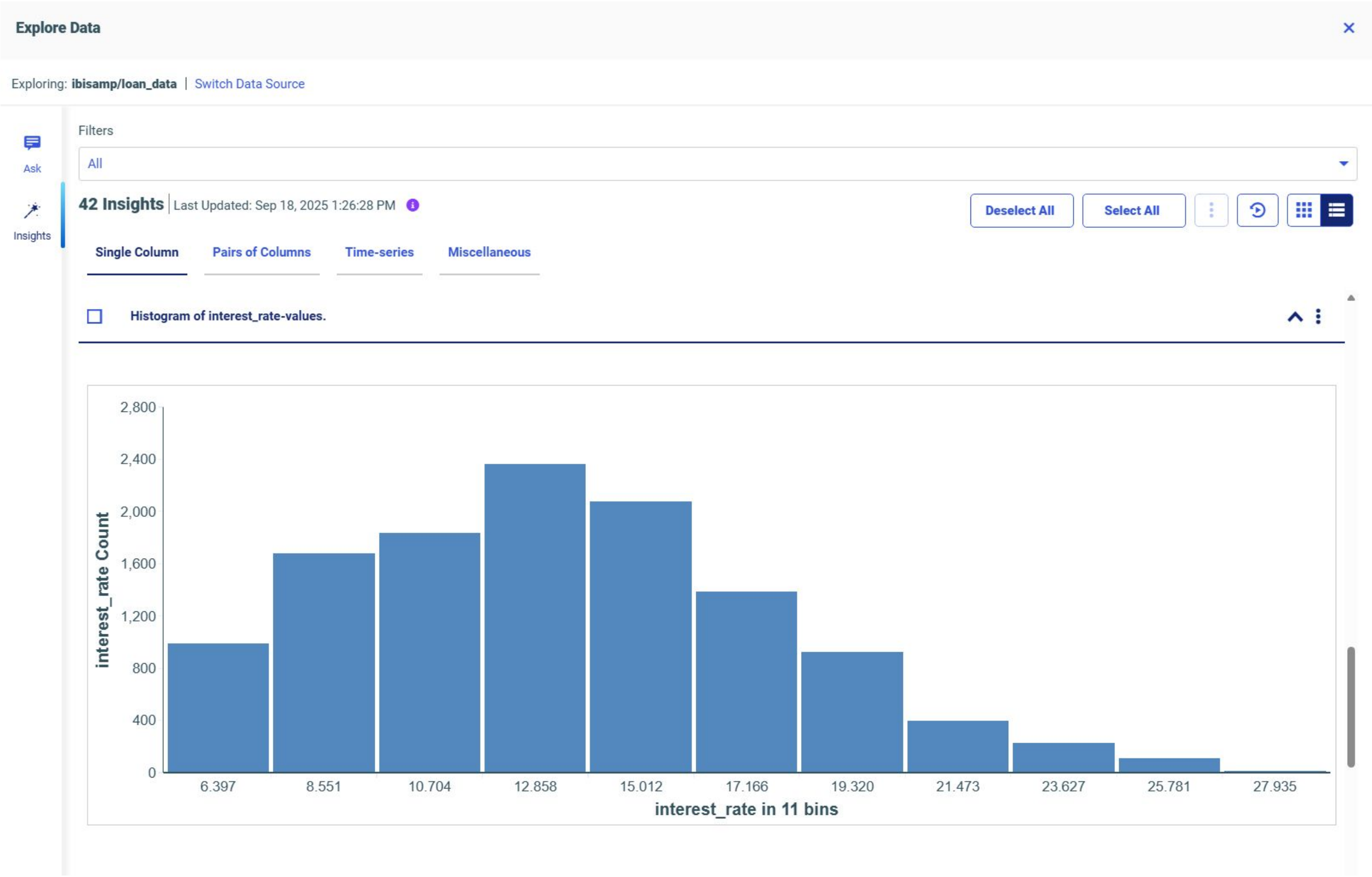
. Interesting results are presented as visualizations.

. Individual visualizations can be moved to the canvas in Designer.

1975 instant photography


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
MORTGAGE and RENT are the most frequent values of home_ownership.
home_ownership
MORTGAGE
RENT
OWN
OTHER


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
debt_consolidation and credit_card are the most frequent values of purpose.
debt_consolidation
credit_card
home_improvement
other
major_purchase
small_business
car
medical
moving
vacation
purpose
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
CNT purpose


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
Histogram of interest_rate-values.
2,800
2,400
2,000
1,600
1,200
800
400
0
interest_rate Count
6.397
8.551
10.704
12.858
15.012
17.166
19.320
21.473
23.627
25.781
27.935
interest_rate in 11 bins

Explore Data

Exploring: **ibisamp/loan_data** | Switch Data Source

Ask

Insights

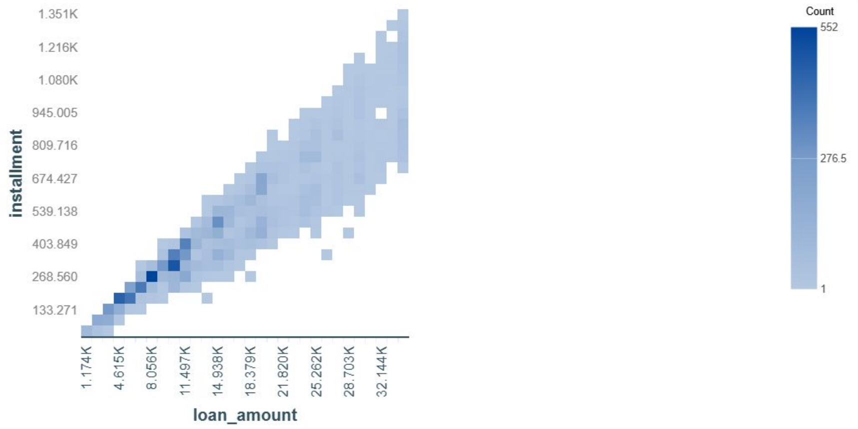
Filters

All

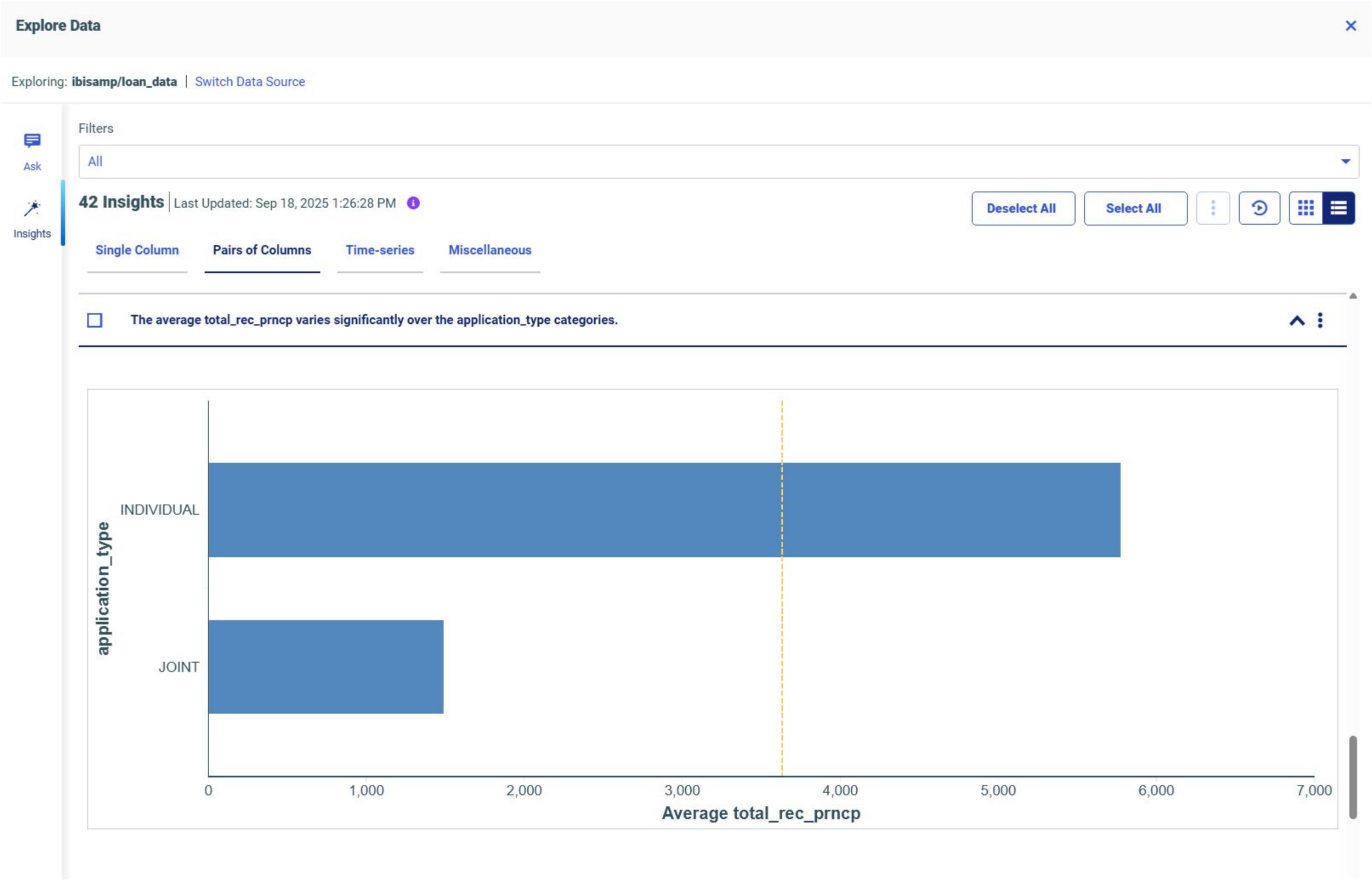
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM

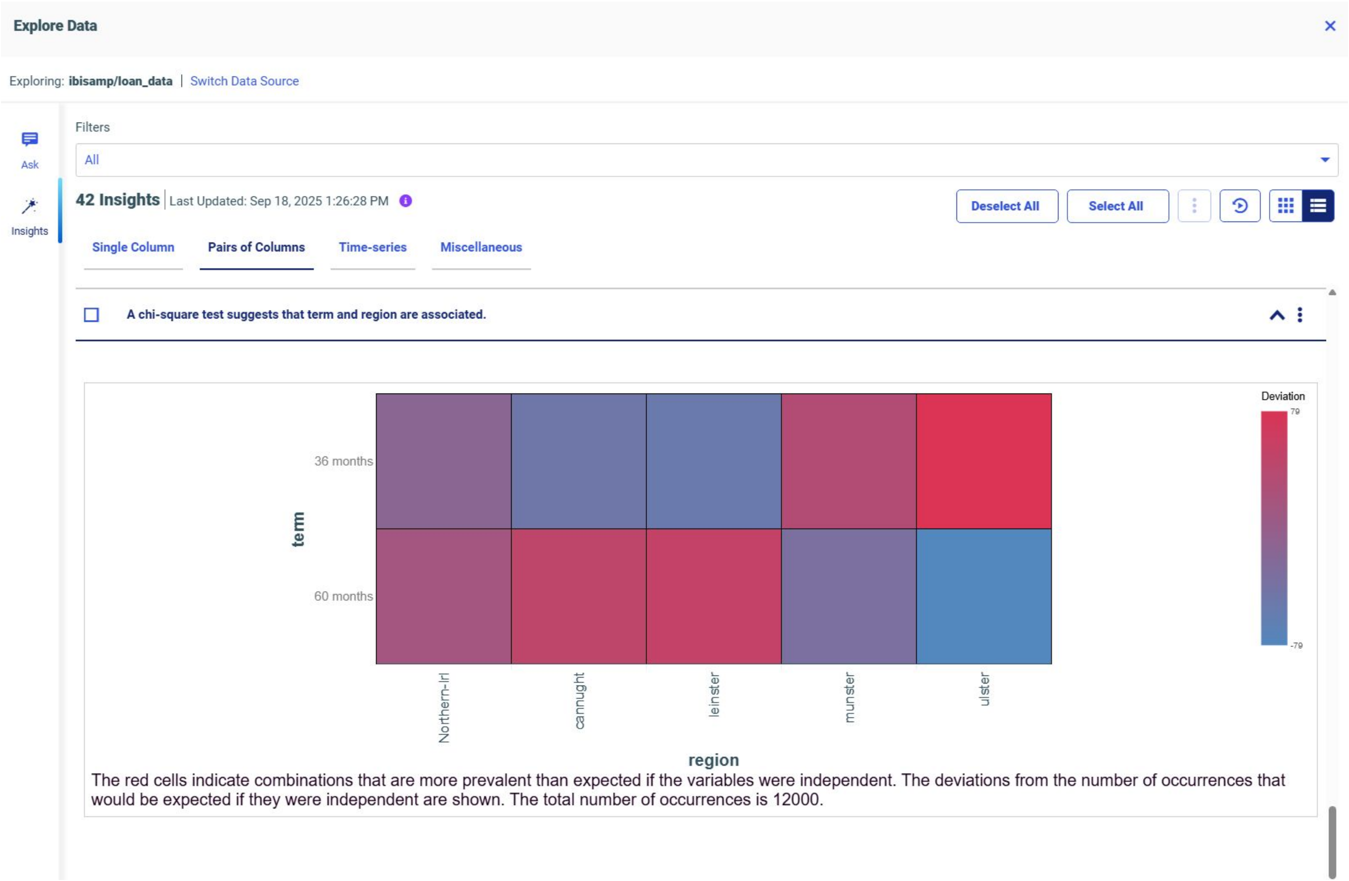
Deselect All | Select All

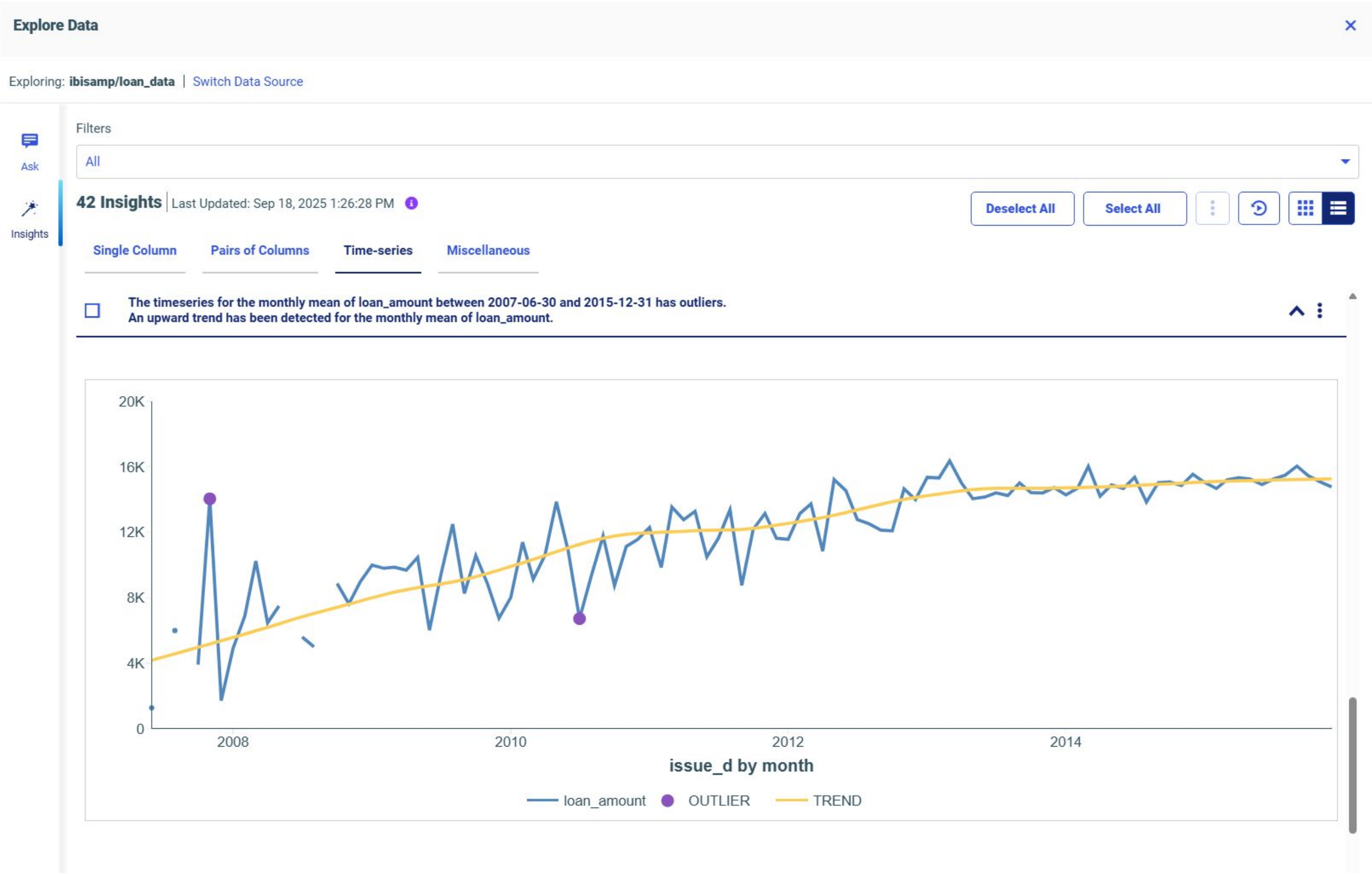
Single Column | **Pairs of Columns** | Time-series | Miscellaneous



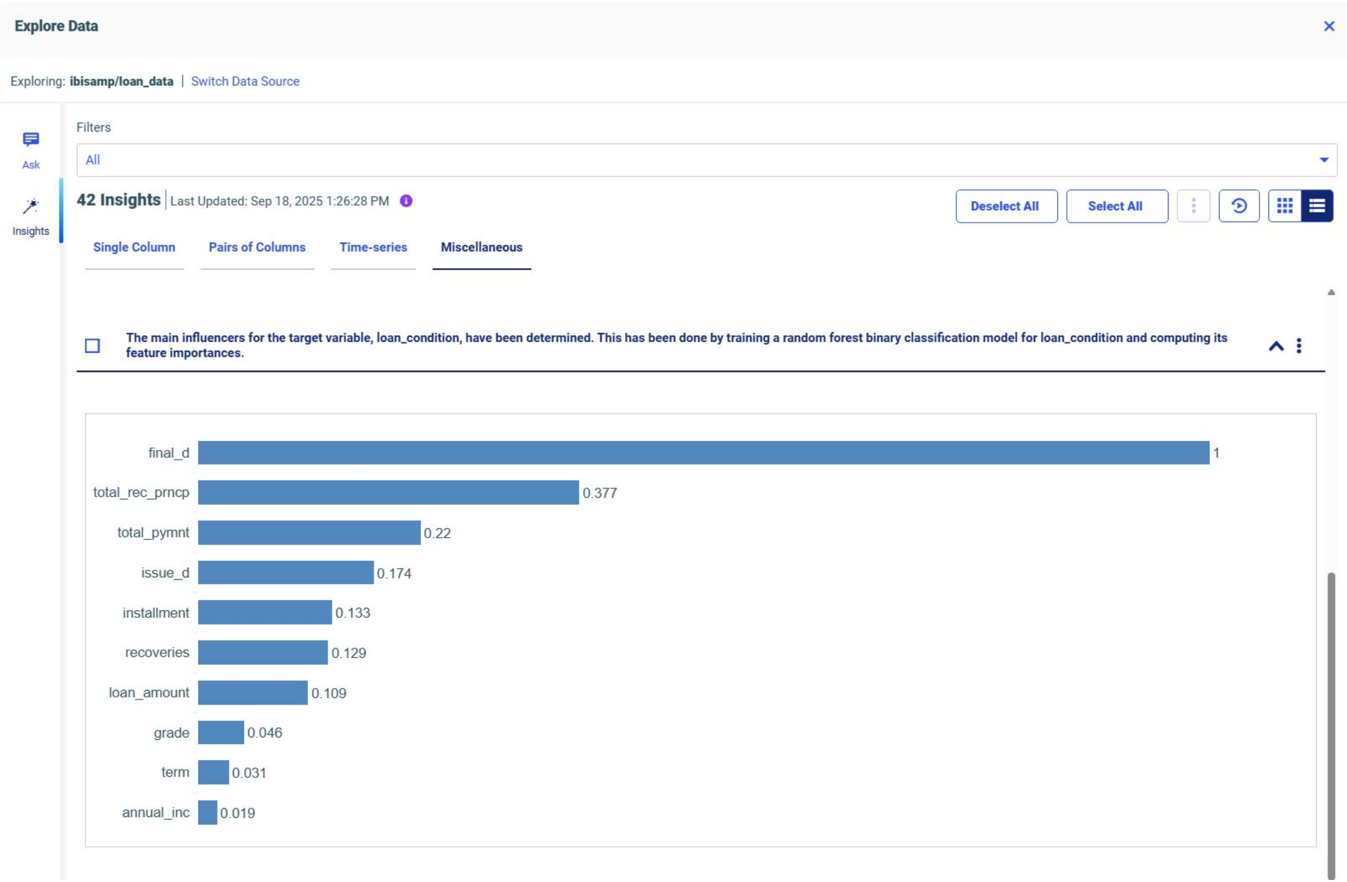
- [ ] There is a medium strength positive correlation between installment and total_pymnt.
- [ ] There is a weak positive correlation between total_pymnt and loan_amount.


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
The average total_rec_prncp varies significantly over the application_type categories.
application_type
INDIVIDUAL
JOINT
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
Average total_rec_prncp


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
A chi-square test suggests that term and region are associated.
Deviation
79
-79
term
36 months
60 months
Northern-Irl
cannught
leinster
munster
ulster
region
The red cells indicate combinations that are more prevalent than expected if the variables were independent. The deviations from the number of occurrences that would be expected if they were independent are shown. The total number of occurrences is 12000.


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
The timeseries for the monthly mean of loan_amount between 2007-06-30 and 2015-12-31 has outliers.
An upward trend has been detected for the monthly mean of loan_amount.
20K
16K
12K
8K
4K
0
2008
2010
2012
2014
issue_d by month
loan_amount
OUTLIER
TREND


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights | Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
The numeric fields have been permuted from the original column-order to emphasize groups of fields that are strongly correlated among themselves, but weakly correlated with the other fields. Such correlated groups lead to a block-diagonal correlation matrix
total_pymnt
total_rec_prncp
installment
loan_amount
annual_inc
dti
interest_rate
recoveries
emp_length_int
Correlation
1.00
0.50
0.00
-0.50
-1.00


Explore Data
Exploring: ibisamp/loan_data | Switch Data Source
Ask
Insights
Filters
All
42 Insights Last Updated: Sep 18, 2025 1:26:28 PM
Deselect All
Select All
Single Column
Pairs of Columns
Time-series
Miscellaneous
The main influencers for the target variable, loan_condition, have been determined. This has been done by training a random forest binary classification model for loan_condition and computing its feature importances.
final_d 1
total_rec_prncp 0.377
total_pymnt 0.22
issue_d 0.174
installment 0.133
recoveries 0.129
loan_amount 0.109
grade 0.046
term 0.031
annual_inc 0.019

Natural Language Query

Ask questions about the data in English

Language model translates English question to SQL SELECT query

SQL SELECT query translated to FOCUS TABLE request

FOCUS TABLE request executed: table output or graph output

50 YEARS OF INNOVATION

Information Builders

Base models

Phi-3 / Phi-4 Small Language Models (SLM) family from Microsoft

Capable small models, trained on high-quality text corpus

. Ideal for local inference

. Ideal for specializing on certain tasks through fine-tuning

WebFOCUS 9.3.6 and earlier

Phi-3 model fine-tuned for English-to-SQL translation, released by Microsoft (3.8 B)

Prompt-engineering, quantization by ibi

WebFOCUS 9.3.7

Phi-4 model fine-tuned and quantized by ibi (3.8 B)

DEMO **NLQ** IN EXPLORE DATA

Remarks on installation of the DSML functionality

On the e-delivery site
https://www.tibco.com/downloads
(stay tuned for a dedicated ibi version, coming soon):

. ibi WebFOCUS Analytics
packages to download + install
(per release, per platform: Windows, Linux)
. Installation guide

50 YEARS OF INNOVATION

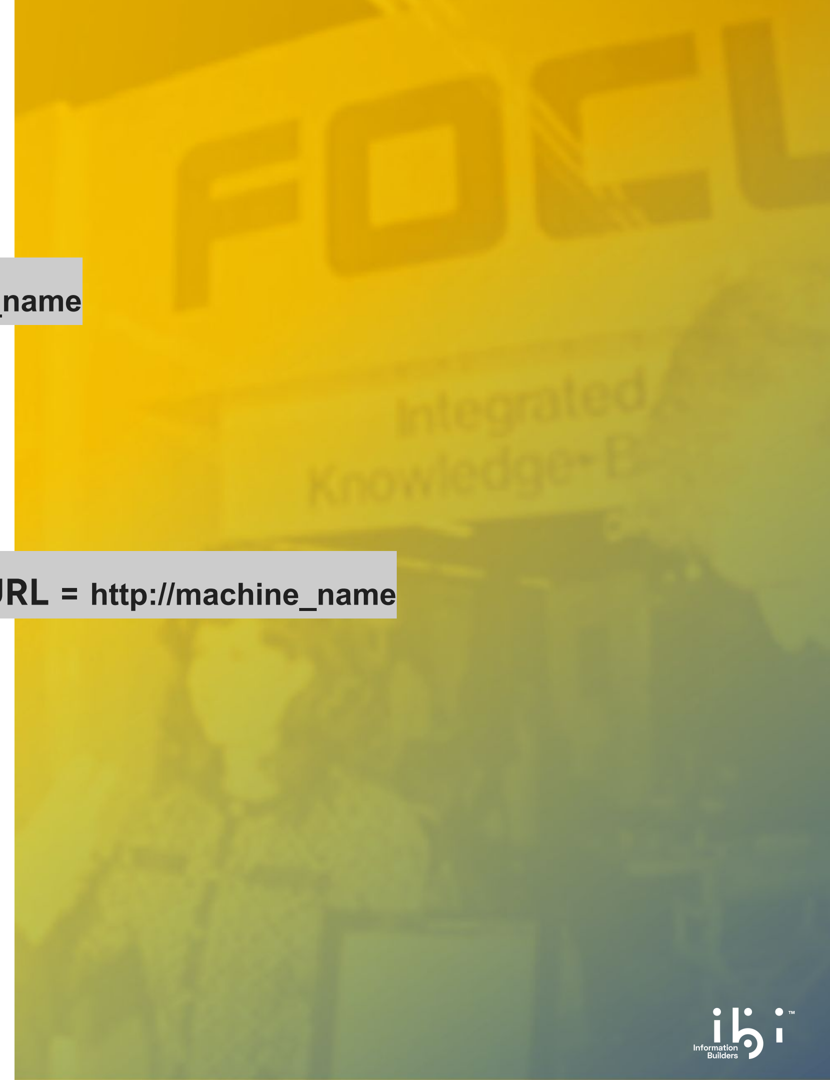
Remarks on installation of the DSML functionality

Configurations

- Workspace - edaserve.cfg
- Administration - admin.cfg
- Shared resources - shared.cfg
- Communication - odin.cfg
- Version - version.cfg
- Server Profile - edasprof.prf
- Trace Profile - ibitrace.fex
- FDS Profile - suprof.prf

Workspace - edaserve.cfg → `pyserv_url = http://machine_name`

Server Profile - edasprof.prf → `SET PYTHON_SERVICE_URL = http://machine_name`



Remarks on installation of the DSML functionality

Why a separate package?

Technically: reporting server makes API calls to the DSML-services from within the reporting server code.

You are not restricted to run the DSML code on the same machine as the reporting server.

Thank you for attending…

Information Builders

AUXILIARY SLIDES

1. RF_CLASSIFY call in focexec
2. Unsupervised Machine Learning
3. Anomaly Detection

Training from within TABLE request

```
TABLE FILE SQLIN
 /*[display]Binary Classification*/
 WRITE
   COMPUTE ATTRITION_CLASSIFY_RF/D7.4 MISSING ON ALL
        TITLE 'Attrition,Random Forest Classification'=CLASSIFY_RF(
'{"pos_label":"Left",
  "scoring":"roc_auc",
  "trees":"100",
  "min_samples_leaf_grid":"1,2,4,8",
  "max_depth_grid":"14",
  "feature_importances":"yes",
  "learning_curve":"yes",
  "kfold":"4",
  "train_ratio":"0.8",
  "test_ratio":"0.2"}',
AGE, YEARS_AT_COMPANY, JOB_ROLE, MONTHLY_INCOME, WORK_LIFE_BALANCE, JOB_SATISFACTION,
PERFORMANCE_RATING, NUMBER_OF_PROMOTIONS, OVERTIME, DISTANCE_FROM_HOME, EDUCATION_LEVEL,
MARITAL_STATUS, NUMBER_OF_DEPENDENTS, JOB_LEVEL, COMPANY_SIZE, COMPANY_TENURE, REMOTE_WORK,
LEADERSHIP_OPPORTUNITIES, INNOVATION_OPPORTUNITIES, COMPANY_REPUTATION, EMPLOYEE_RECOGNITION,
ATTRITION );
```

Random Forest Classification

Parameters and hyperparameters (defaults)

X

y

General structure of problems addressable with ML

Unsupervised Machine Learning

Anomaly Detection, Clustering

. **Find structure in the data without specifying a target column: outliers, clusters**

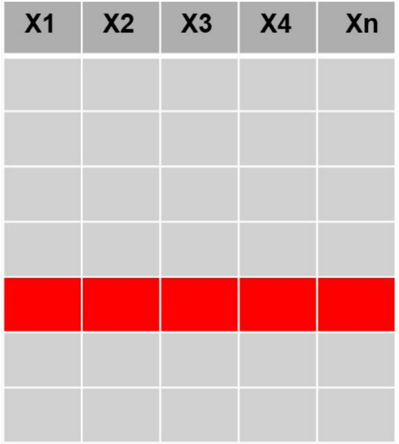
. **Model F: outlier_score = F(X) or cluster_number = F(X)**

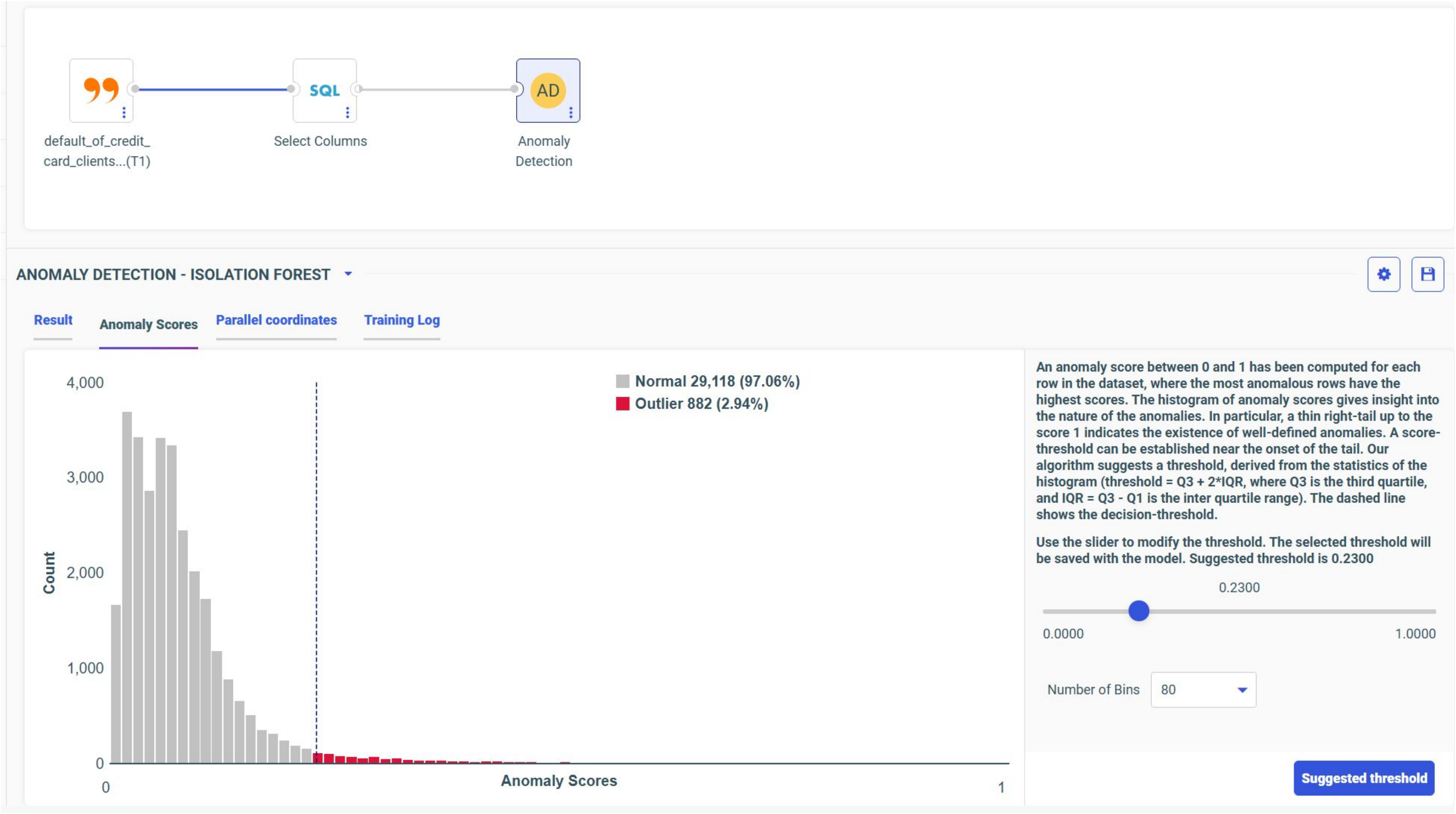
Clustering

groups of similar instances

Anomaly Detection

minority of outlier instances





default_of_credit_card_clients...(T1)
SQL
Select Columns
AD
Anomaly Detection
ANOMALY DETECTION - ISOLATION FOREST
Result
Anomaly Scores
Parallel coordinates
Training Log
Normal 29,118 (97.06%)
Outlier 882 (2.94%)
4,000
3,000
2,000
1,000
0
Count
0
Anomaly Scores
1
An anomaly score between 0 and 1 has been computed for each row in the dataset, where the most anomalous rows have the highest scores. The histogram of anomaly scores gives insight into the nature of the anomalies. In particular, a thin right-tail up to the score 1 indicates the existence of well-defined anomalies. A score-threshold can be established near the onset of the tail. Our algorithm suggests a threshold, derived from the statistics of the histogram (threshold = Q3 + 2*IQR, where Q3 is the third quartile, and IQR = Q3 - Q1 is the inter quartile range). The dashed line shows the decision-threshold.
Use the slider to modify the threshold. The selected threshold will be saved with the model. Suggested threshold is 0.2300
0.2300
0.0000
1.0000
Number of Bins
80
Suggested threshold