Hands-On Tutorial Human Resources employee Churn

Author: Herve KAUFFMANN

This guide will walk us through the process of forecasting cost for a Human Resources manager Identify high performing at risk employees and to reduce employee turnover.

EMPLOYEE DATASET - DESCRIPTION OF THE VARIABLES

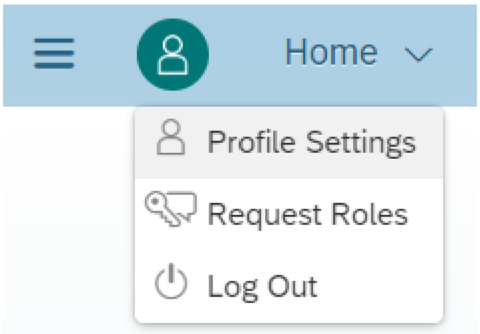
Name in the dataset	Description	Value	Storage	Type	Role
Employee ID	Employee ID.	Predefined ID	Integer	Continuous	Unique identifier
NAME	Employee Last Name		String	Nominal	Explanatory variable
FIRST_NAME	Employee Birth first name		String	Nominal	Explanatory variable
GENDER	Female or Mzle	F, M	Predefined list	Nominal	Explanatory variable
MANAGER	Manager of the employee	Employee ID of the manager	Integer	Continuous	Explanatory variable
EMPLOYEE_TYPE	Internal or External	EMP = Internal	String	Nominal	Explanatory variable
DAPARTMENT	Department of the employee	Departmen t ID	String	Nominal	Explanatory variable
DPT_CHANGE_FLAG		R, C, E	Predefined list	Nominal	Explanatory variable
JOB	Job description	Predefined list	String	Nominal	Explanatory variable
STATUS		A	Predefined list	Nominal	Explanatory variable
COMPANY	Subsidiary where the employee works	Subsidiary ID	String	Nominal	Explanatory variable
SITE	Place where the employee works	Site ID	String	Nominal	Explanatory variable
PERMANENT	Type of contract	E, R	Predefined list	Nominal	Explanatory variable
EMPLOYEE_CLASS	Type of contract	INT, IMP, NULL	Predefined list	Nominal	Explanatory variable
FULL_TIME	Type of contract	E, F, P	Predefined list	Nominal	Explanatory variable
EMPLOYEE_LEVEL	Type of Contract	A, B, C, T, Null	Predefined list	Nominal	Explanatory variable
HANDICAP	Type of contract	MOTD, MOTL, VISU, NULL	Predefined list	Nominal	Explanatory variable
CITIZENSHIP	Country of the employee	COUNTRY CODE	Predefined list	Nominal	Explanatory variable
AGE	Age of the employee	In years	Integer	Continuous	Explanatory variable

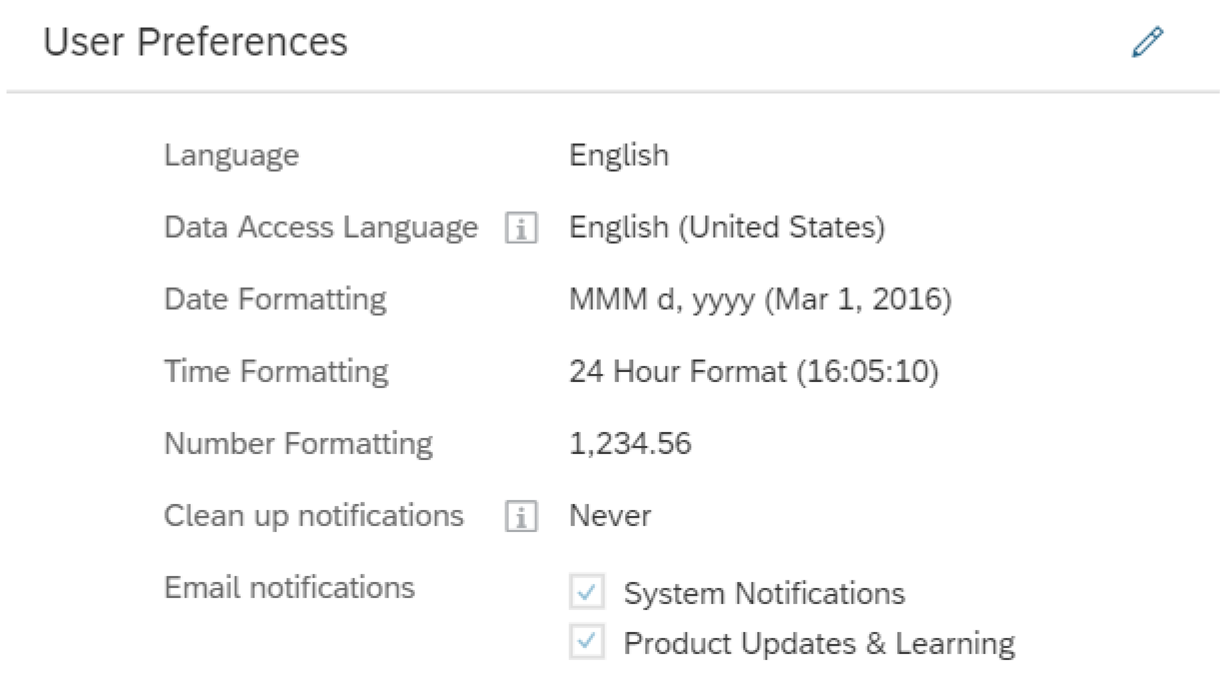
CONTRACT_TENURE	Number of years of the current contract	In Years	Integer	Continuous	Explanatory variable
EMPLOYEE_TENURE	Number of years of the employee in the company	In years	Integer	Continuous	Explanatory variable
SUM_BONUS_UNEXPECTED_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_WELCOME_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_CHALLENGE_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_MISC_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_EXC_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_LANGUAGE_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_SHARING_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_OBJECTIVE_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_YIELD_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_TECHNICAL_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_TOTAL_3Mago	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_UNEXPECTED	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_WELCOME	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_CHALLENGE	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_MISC	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_EXC	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_LANGUAGE	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_SHARING	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_OBJECTIVE	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_YIELD	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_TECHNICAL	Special Bonus	In €uros	Number	Continuous	Explanatory variable
SUM_BONUS_TOTAL	Special Bonus	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_CHALLENGE	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_LANGUAGE	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_MISC	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_OBJECTIVE	Index	In €uros	Number	Continuous	Explanatory variable

EVOLUTION_BONUS_SHARING	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_TECHNICAL	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_TOTAL	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_UNEXPECTED	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_WELCOME	Index	In €uros	Number	Continuous	Explanatory variable
EVOLUTION_BONUS_YIELD	Index	In €uros	Number	Continuous	Explanatory variable
Target_Churn	Employee still in the company	0=YES, 1=NO	String	Nominal	Target Variable

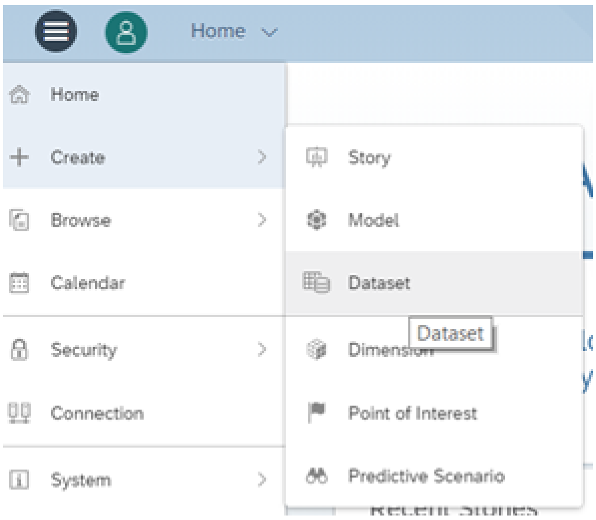
First log on to a SAP analytics Cloud instance.

Before we start, have a look at you profile setting and make sure the number formatting is set to “1,234.56”.

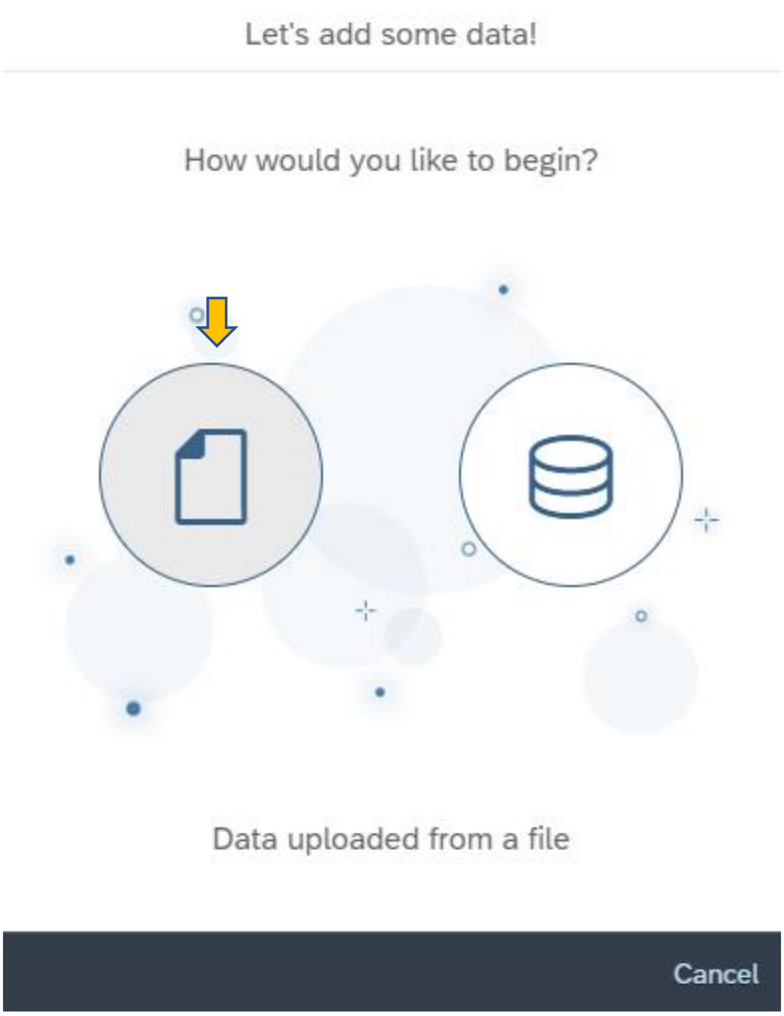




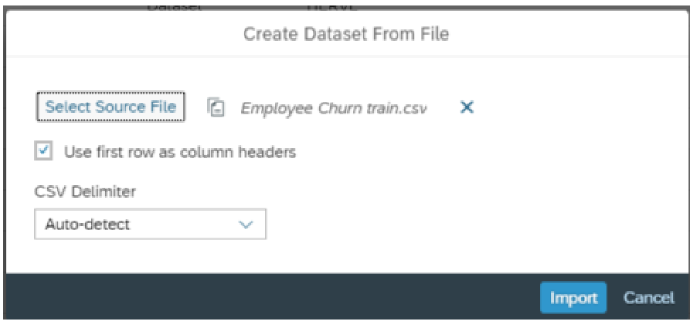
After the logon the dataset needs to be uploaded. To do this we click on the menu on the top left, select “Create” and click on “Dataset”.



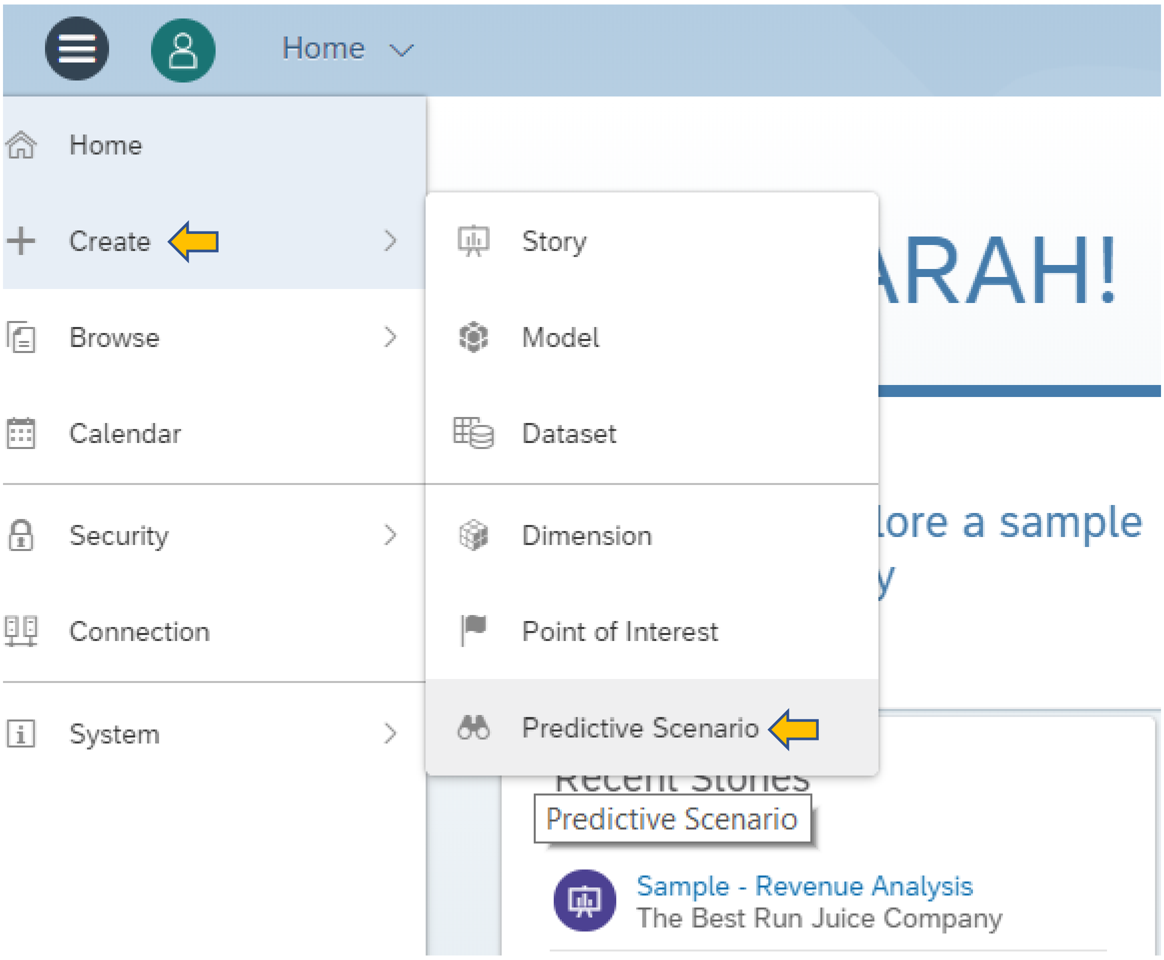
On the Pop Up, we select “Data uploaded from a file”.



We select the source file “Employee Churn train.csv”, click “Import” and then “Ok”.



Now that we have uploaded the data set we can start to build our predictive scenario. We select “Create” and then “Predictive Scenario” on the menu.



A predictive scenario set of use cases with common characteristics. SAP Analytic Cloud's Smart Predict currently offers 3 predictive Scenarios:

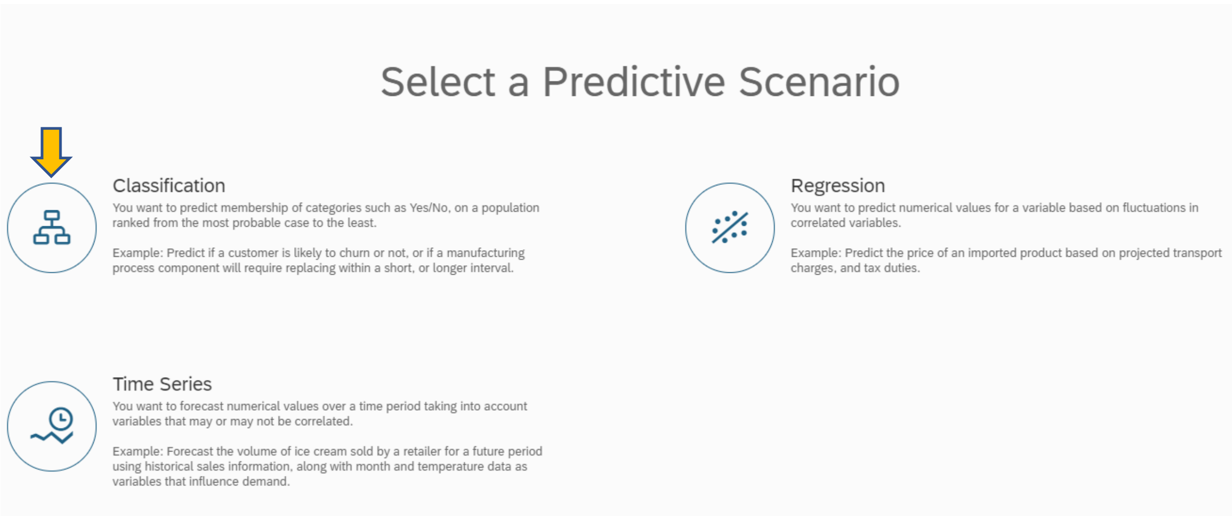
- **Classification scenarios** predict the value of a (target) variable that can only have two values like yes and no or 0 and 1. Examples for classification scenarios are
  - customer churn with the target variable predicting whether a customer will leave or not
  - Propensity to buy with the target variable predicting whether a customer will buy a product offered to him or not
  - Fraud with the target variable indicating whether a transaction or claim was fraudulent or not
- **Regression scenarios** predict the numerical value of a target variable depending on variables describing it. Example for regression scenarios are the prediction of
  - The number of customers visiting a shop during lunch time
  - The revenue of a customer in the next quarter
  - The sales price of a used cars
- **Time Series scenarios** predict the value of a variable over time taking into account further descriptive variables. Examples of time series scenarios are the prediction of
  - Revenue for a product line over the next few quarters
  - The number of bicycles hired in a city over the next few days
  - Travel expenses in the next few months

The user now must follow 3 simple steps:

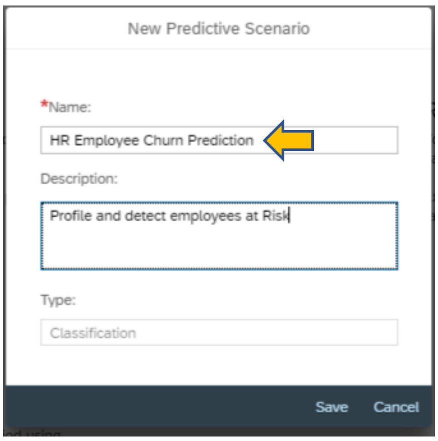
1. Choose the predictive scenario that matches his use case.
2. Train the model with historic data, i.e. use a data set where sales figures are known. The statistical algorithm will "learn" from this data set, i.e. find trends, seasonal variations and fluctuations that characterize the sale of a certain product. There should be enough (3-5 years) data available to learn from.
3. Apply the model to a new data set, i.e. forecast sales for a given period of times. The statistical algorithm will apply the patterns learnt in the previous step to the new data and predict sales for the chosen number of time periods.

The variable that contains churners in the learning phase and is predicted in the application phase is called the target variable.

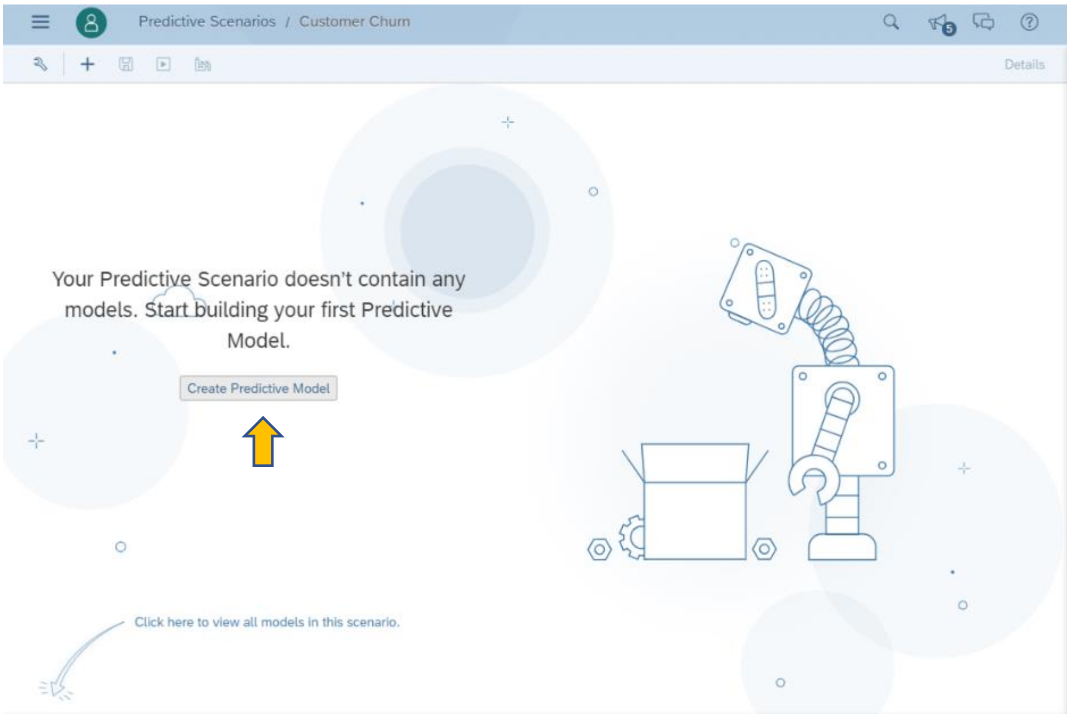
The following screen shot shows the three options classification, regression and time series. Under each option there is a description to make it easier for the user to select the right scenario for each use case. In this exercise, we want to detect which employees of the company are at risk. Based on the descriptions of predictive scenario types, you can see that a classification will be able to address our needs. So, we select it.



On the Pop Up we give the model a name, e.g. "HR Employee Churn Prediction". We enter the Business Question as "T&E Cost Prediction for 2018".

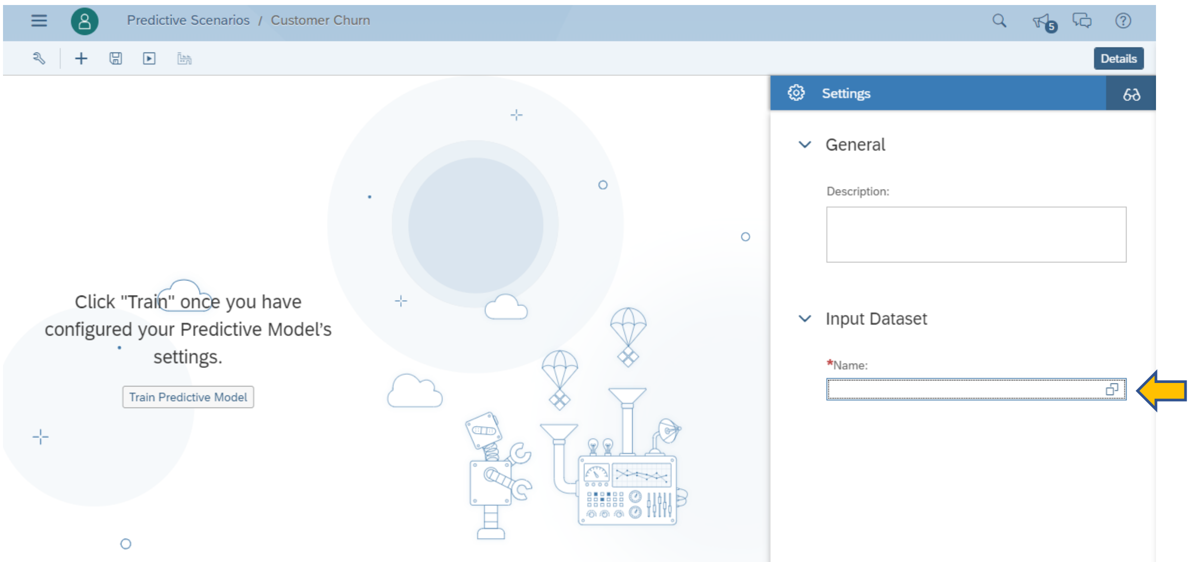


Now we can create our Predictive Model.

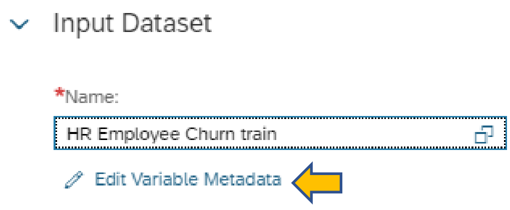


We will need to select an input dataset for our model. The input data set contains historical data that we use to train the predictive model.

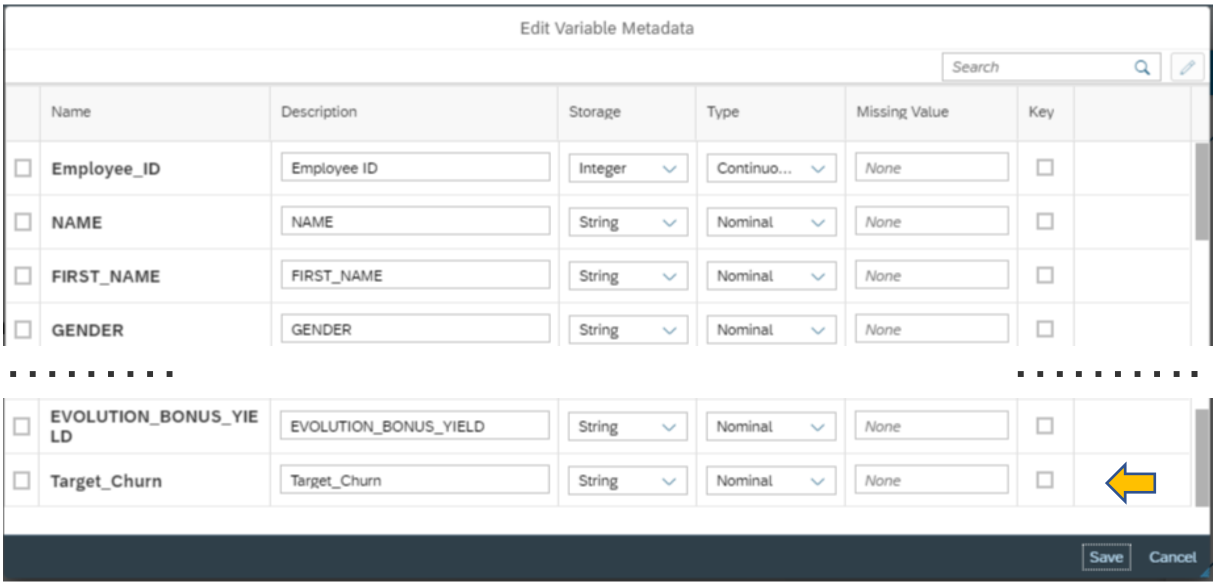
Select "HR Employee Churn train" from the folder.



After selecting the input data let's have a look at the variable metadata. We click on "Edit variable Metadata" directly below the field where we selected the input data and check that all data types of variables were correctly identified.



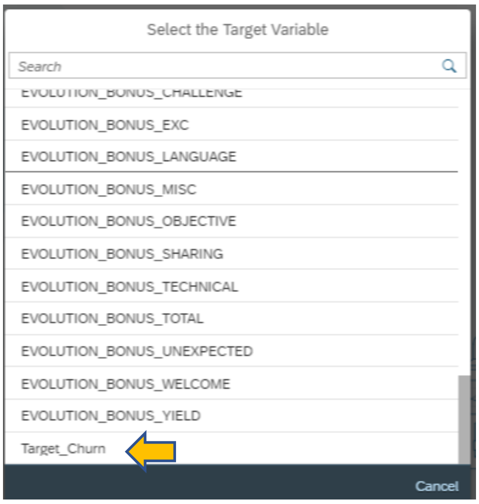
Please check that all data types of the variables where recognized correctly as you see in this screenshot.



The Target variable to predict is "Target_Churn". Classification scenarios predict the value of a variable (the target variable) that can only have two values like yes and no or 0 and 1.

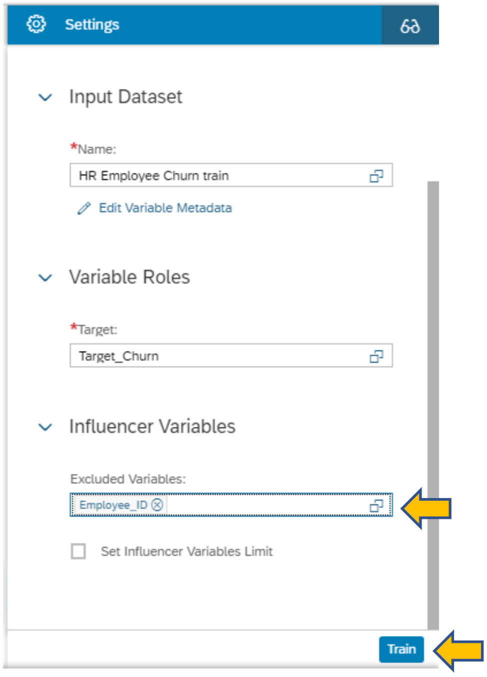
The Target variable in this scenario contains the critical employee information we need to predict the behavior of all active employees.

Target_Churn = 0 means that the employee is active in the company. Target_Churn = 1 means that the employee has left company. This variable has been created prior to build the predictive scenario. The rules to build a target variable must be very precise and can be complex.



Variables that have no influence on the target can be excluded from the modeling process. Excluding variable can speed up the execution process but keeping them does not interfere with the modelling process. IDs are typical variables to exclude.

However, you must exclude variables that are directly related to the target variables such as transformations of the target variables and variables that contain the same information as the target variable indirectly. For example, if a dataset contains two fields that contain the cots number maybe just in different currencies you need to exclude one variable.



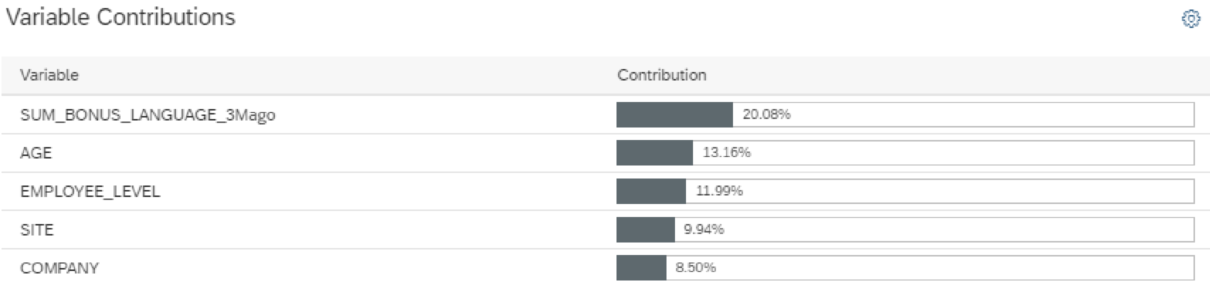
Click “train predictive Model” and when prompted to save, we click ok.

Be patient since this might take a couple of minutes.

After the model was trained, we select version one. We see two performance indicators that describe the quality of the model. The Predictive Power indicates the proportion of information contained in the target variable that the model and the explanatory variables can explain.

The Prediction Confidence shows the robustness. It signifies the capacity of the model to achieve the same performance when it is applied to a new data set exhibiting the same characteristics as the training data set.

The chart below shows the variables that the model generation process identified as relevant on the left and orders them by their impact on the target variable.

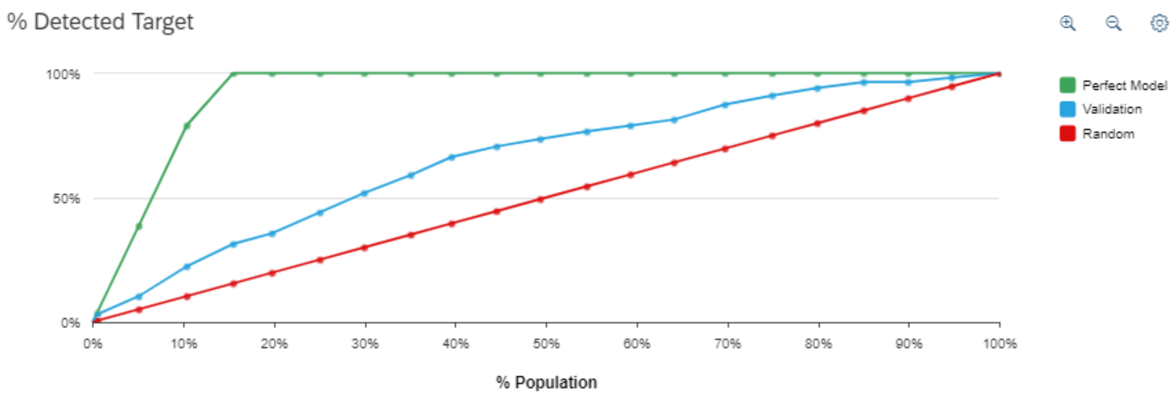


On the bottom of the screen we see the performance curve. The X axis shows a percentage of the initial population; the Y axis represents the percentage of positive targets the classification algorithm detected.

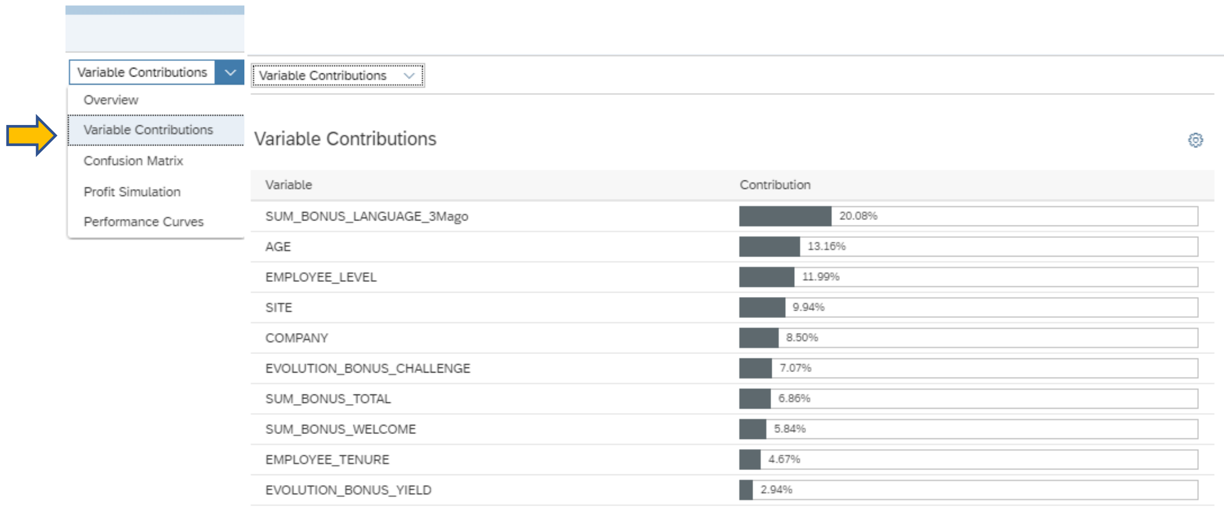
The green curve shows the maximum possible percentage of detected target, obtained by a perfect model. For example, if 25% of the employees has the target category “Has churned”, then the best model would correctly classify all 25% of the employee that have Churned within 25% of the population.

The red curve shows the minimum percentage of detected target, obtained by a random model. By randomly taking 10% of the population, you would identify 10% of these employees.

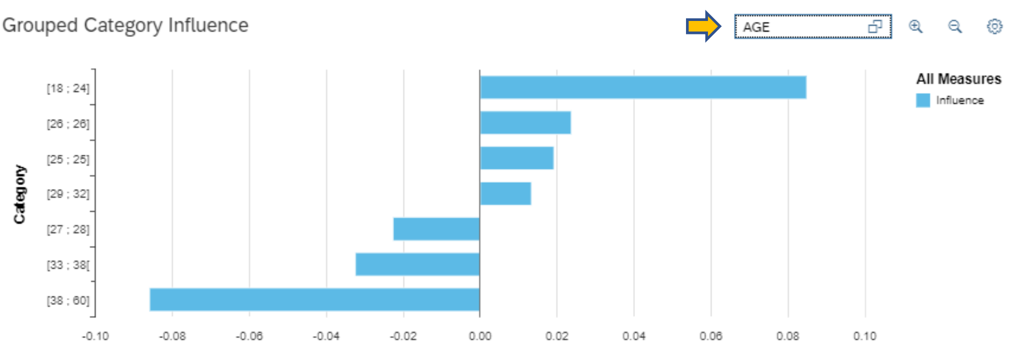
The blue curve shows the percentage of detected target, obtained by the Smart Predict model. For example, if we take 10% of the population we detect roughly 40% of these employees. The closer the blue line gets to the green line the better is the model. The larger the distance between the red and the blue line the bigger is the lift of the model.



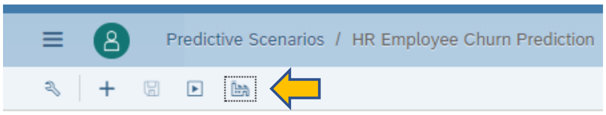
We want now to understand the influencing factors better. So, we have a look at the other options besides “Overview”. We select “Variable Contributions” on the top left to better understand the impact and the detailed influence the contributing variables have on the target.



We want now to understand how a variable can influence the risk of churn. At the bottom of the screen we can select the variable we want to investigate. For example, youngers employees are more likely to churn ([18-24]).



We are now convinced of the quality of the Smart Predict Classification model, and we want to apply it. We click on the little factory icon on the top left.

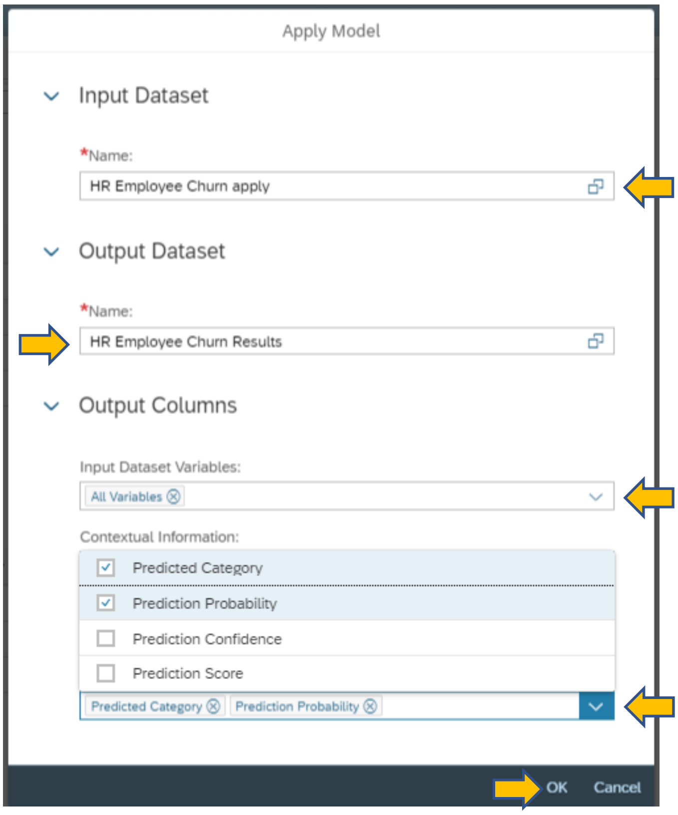


We choose the dataset you want to score (the input dataset); in our example, we select “HR Employee Churn Apply” from the list; it contains all the active employees from the company.

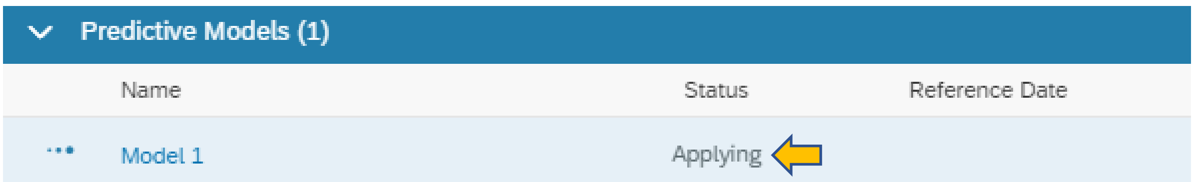
We also choose in which folder we want to store the results and give a name to the dataset generated by the applying (the output dataset). For example, “HR Employee churn Result”.

We must also decide what information we want in the output dataset. Here we have chosen to insert all variables from the input dataset + Apply date + Predicted Category + Probability to churn.

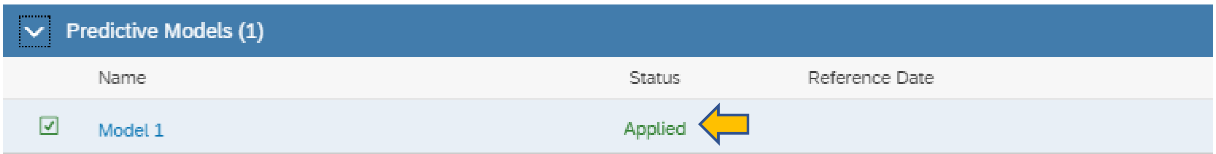
When this is done **click OK to apply**.



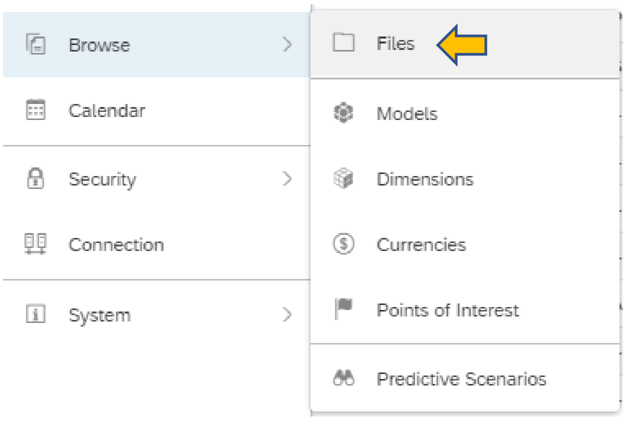
Be patient since this might take a couple of minutes.



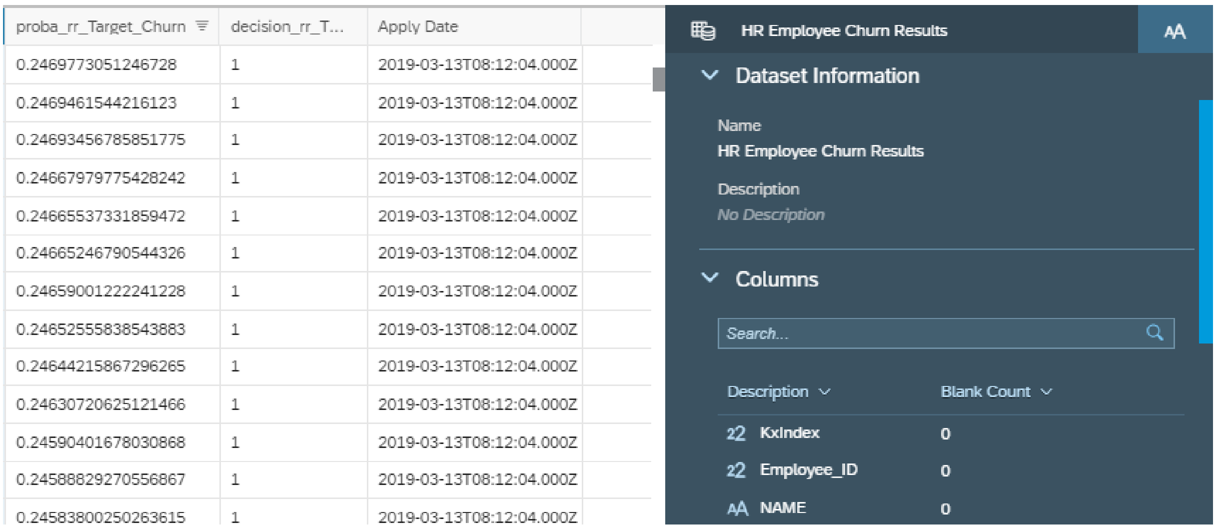
When this is done the status change from Applying to Applied.



We navigates to the folder where he saved the output data set to view the results.



On the far right we see the column “decision_rr_T…” that contains the prediction whether the employee will churn or not based on the probability in column “proba_rr_Target_Churn” assigned by the algorithm of the Smart Predict model. The higher the probability the more likely it is that the employee will churn, the lower it is the more likely it is that it will be stay in the company. Thanks to the “Apply Date” we know when the risk to churn has been calculates.



To Continue open next tutorial : Hands-On BI Story Tutorial Employee Churn