PLEASE LEAVE THIS SHEET IN THE ENVELOPE WITH THE NUMBER VISIBLE THROUGH THE WINDOW

C11 -Managing Knowledge and Artificial Intelligence

G00 *G00*

C11

Content

- 1 (this) instruction sheet **(Back in the envelope)**
- Log file "MIS_C11_STUDENT.zip" can be downloaded from Moodle
- 1 TABLE with variables of interest **(Back in the envelope)**
- 1 ANSWER sheet with Persona and marketing campaign recommendation **(Back in the envelope)**

Instructions

If you have not done it yet, follow the setup instructions on Moodle.

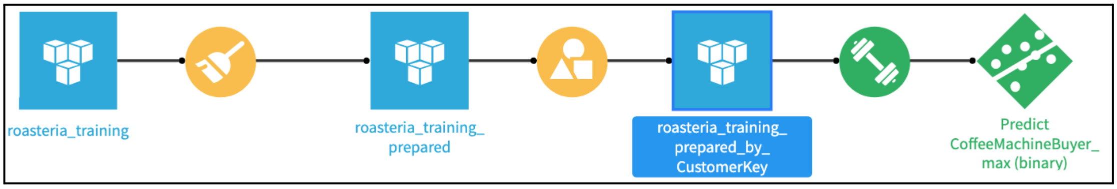
4- Build a ML classifier

a. Now that we have prepared the dataset it is time to train our first ML model: a classifier! We want to learn whether features of our dataset can be used to predict whether a client will buy a coffee machine or not. Select the last dataset and click on the LAB button on the right hand side. Then select "AutoML Prediction". Our target feature will be "CoffeeMachineBuyer_max". Let's stick with an AutoML prototype and with "Quick Prototypes" and click on the "Create" button. Before training your model, review the design. Go to the advanced options, "Runtime environment". Decrease the "Max concurrency" setting to "1". Save the model. Now you can Train your model. You can learn how to complete a similar procedure following this tutorial tinyurl.com/dataikuclass.

b. You will notice that Dataiku will compare two ML algorithms: Random Forest and Logistic Regression. Look at the comparison and decide which model to keep going forward. Add the details of your choice in the "TABLE variables of interest sheet".

c. Get in the details of the model that you have chosen and study the "Feature importance" view. This provides an overview of the features of the model. Report the top-5 features in the "TABLE variables of interest sheet". For the "Value of Reference" study the "Feature Dependance" view and choose a reference point(s) for each feature which trigger(s) positive Shapley values for the model.

d. Deploy the model to your project.


roasteria_training
roasteria_training_ prepared
roasteria_training_ prepared_by_ CustomerKey
Predict CoffeeMachineBuyer_ max (binary)

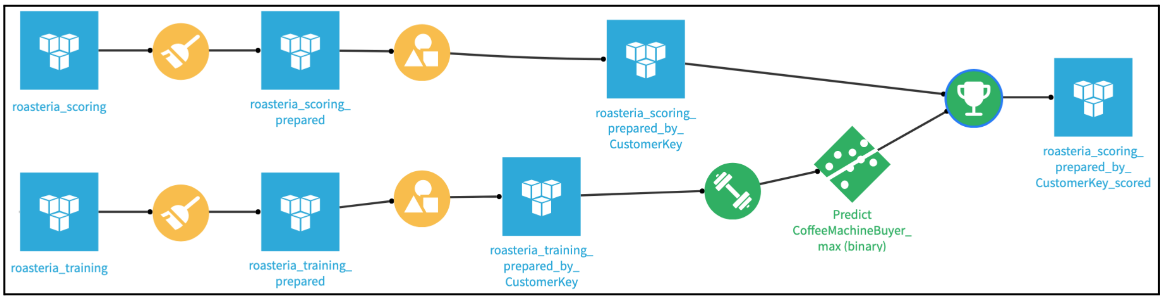
5- Score new data

a. Once you have a model deployed to the Flow, you can use it to generate predictions on new, unseen data. You can learn how to complete a similar procedure following this tutorial tinyurl.com/dataikuscor.

b. After step 3, you should have a dataset with the exact same set of features as the training data. The only difference is that, in the data ready to be scored, the values for the CoffeeMachineBuyer column, which is the target variable, are entirely missing, as we do not know which new customers will buy a coffee machine. Let's use the model we previously built to generate predictions of whether the new customers will buy a coffee machine. From the Flow, select the deployed prediction model, and add a

"Score" recipe from the Actions sidebar. Choose "roasteria_scoring_prepared_by_CustomerKey" as the input dataset. Click Create Recipe.

c. Leave the default recipe settings, click Run, and then return to the Flow.

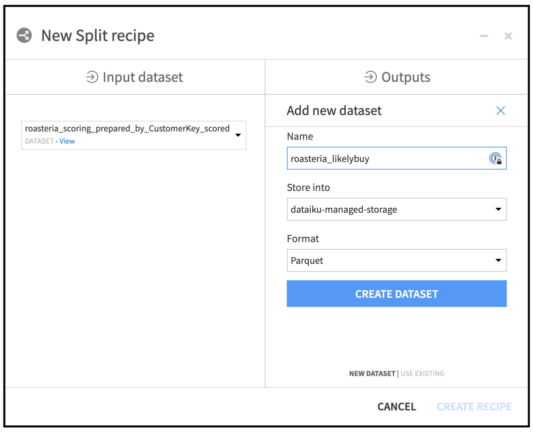

roasteria_scoring
roasteria_scoring_prepared
roasteria_scoring_prepared_by_CustomerKey
roasteria_scoring_prepared_by_CustomerKey_scored
roasteria_training
roasteria_training_prepared
roasteria_training_prepared_by_CustomerKey
Predict CoffeeMachineBuyer_max (binary)

d. Inspect the scored data. There are three new columns appended to the end:
   a. proba_0 is the probability a customer will **not** buy a coffee machine.
   b. proba_1 is the probability a customer will buy a coffee machine.
   c. prediction is the model's prediction of whether the customer will buy (i.e., 0 = no; 1 = yes).

e. Use the Analyze tool on the prediction column to see what percentage of new customers the model expects to buy a coffee machine.

6- Analyze the characteristics of customers who will likely buy and those who might not buy

a. Now that you have scored new customers we can use this information to split the participants in two groups: those who will **likely buy** a new coffee machine, based on their characteristics, and those who **might not buy** a machine. There are several ways to possibly do this. We advise you to split the dataset in two using the "Split" recipe. Create two new dataset: "roasteria_likelybuy" and "roasteria_mightnotbuy". Map values of a single column to the output datasets. Select the column named "prediction" and assign rows with value 1 to "roasteria_likelybuy". Save and run.



New Split recipe

Input dataset

roasteria_scoring_prepared_by_CustomerKey_scored
DATASET - View

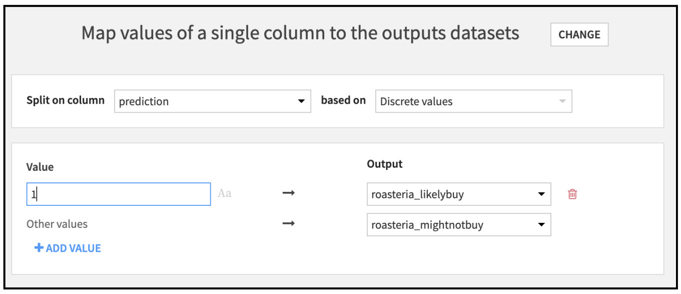
Outputs

Add new dataset

Name: roasteria_likelybuy

Store into: dataiku-managed-storage

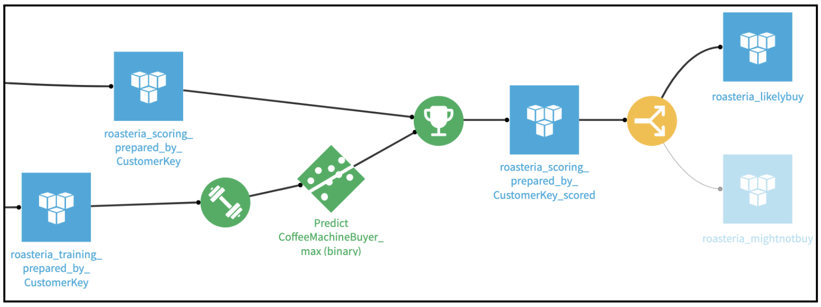
Format: Parquet



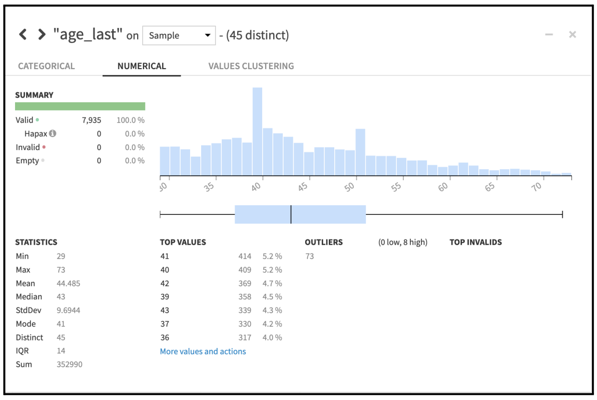
Map values of a single column to the outputs datasets

Split on column: prediction — based on: Discrete values

Value		Output
1	→	roasteria_likelybuy
Other values	→	roasteria_mightnotbuy


roasteria_scoring_prepared_by_CustomerKey
roasteria_training_prepared_by_CustomerKey
Predict CoffeeMachineBuyer_max (binary)
roasteria_scoring_prepared_by_CustomerKey_scored
roasteria_likelybuy
roasteria_mightnotbuy

b. Once the split has been completed, you can open each dataset and analyze the variables of interest, clicking on "Analyze" command from the drop down menu.



"age_last" on Sample - (45 distinct)

CATEGORICAL | NUMERICAL | VALUES CLUSTERING

SUMMARY

Valid	7,935	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %

STATISTICS

Min	29
Max	73
Mean	44.485
Median	43
StdDev	9.6944
Mode	41
Distinct	45
IQR	14
Sum	352990

TOP VALUES

41	414	5.2 %
40	409	5.2 %
42	369	4.7 %
39	358	4.5 %
43	339	4.3 %
37	330	4.2 %
36	317	4.0 %

More values and actions

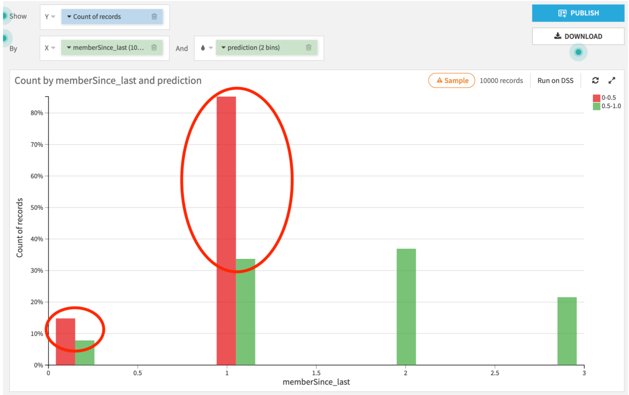
OUTLIERS (0 low, 8 high)

73

TOP INVALIDS

c. (alternative method) Alternatively, you can open the last "...scored" dataset from the Flow and select "Charts" from the top menu. You might want to pick a "Vertical bars" chart. On the x-axis, select the variable of preference (e.g., MemberSince_last). On the y-axis, you can select "Count of records" and set as aggregation a "Percentage scale". To differentiate between buyers and non-buyers, you can assign the variable "prediction" to the color split, and set the boundary to 2 bins.

d. (alternative method) As you can see, non-buyers –here in red– are clients that have created their membership profile less than a year ago.



Show Y: Count of records

By X: memberSince_last (10... And prediction (2 bins)

Count by memberSince_last and prediction

e. You can repeat this analysis for all the variables of interest. Additionally, you can also add these charts to the project dashboard, for future references.

7- Build the Persona of the clients and define your marketing campaign

a. It is now time to wrap up the analysis and define the prototypical profile of your customers. Fill the "ANSWER Persona and marketing recommendation" document by studying the different profile variables.

b. Once you have identified the characteristics of the profile of the clients, think how you might have these customers buy a coffee machine. Would you offer a discount to a particular sub-population of your customer? Would you do more advertisements? Would you set up a promotion? Motivate your choice based on data and on the profile of the clients.

c. Lastly, explain how you would deploy your marketing campaign. How would you make your customer aware of this marketing campaign?

—