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

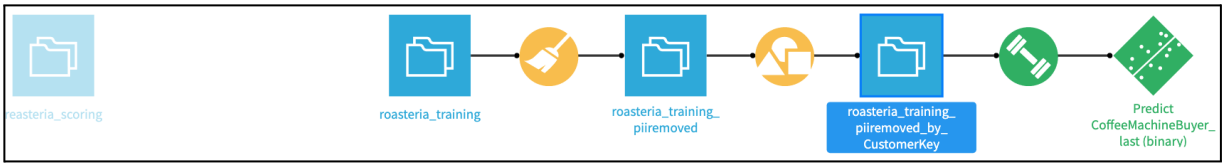
3- Build a ML classifier

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_last". Let's stick with an AutoML prototype and click on the Train button. 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 "Variable importance" view. This provides an overview of the features of the model. Report the top-5 features in the "TABLE variables of interest sheet".

d. Deploy the model to your project.

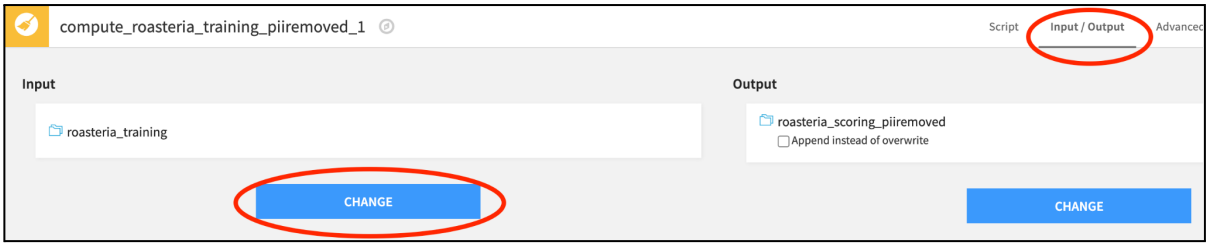


4- Score new data

a. Once you have a model deployed to the Flow, you can use it to generate predictions on new, unseen data. This new data, however, first needs to be prepared in a certain way because the deployed model expects a dataset with the exact same set of features as the training data. You can learn how to complete a similar procedure following this tutorial tinyurl.com/dataikuscor.

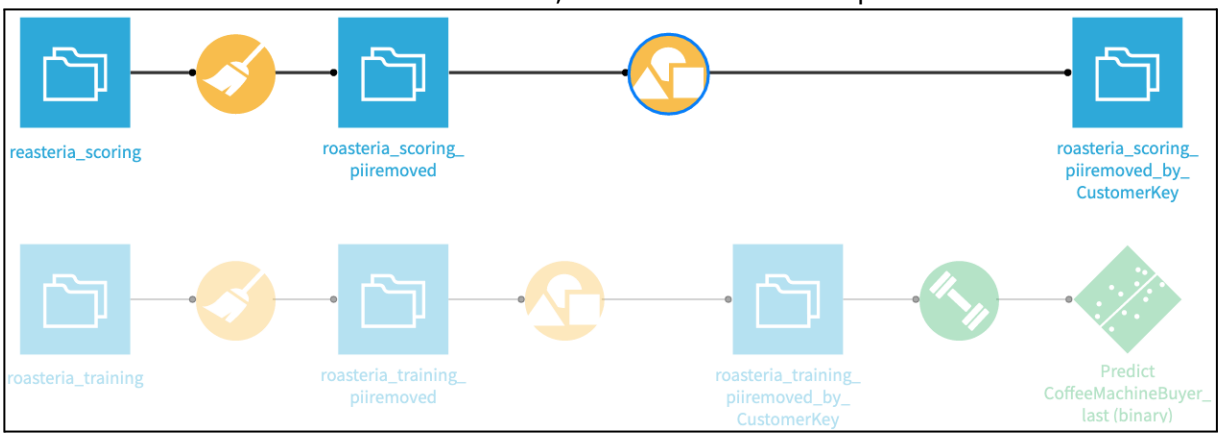
b. Copy the recipes to prepare the scoring dataset. From the Flow, select both the Prepare and Group recipes (On a Mac, hold Shift to select multiple objects). From the Actions sidebar, choose Copy. From the Flow View menu at the bottom left, choose Copy To. Rename "roasteria_training_piiremoved" "roasteria_scoring_piiremoved" and "roasteria_training_piiremoved_by_…" as "roasteria_scoring_piiremoved_by…". Once done select Copy.

c. The Flow now looks like a mess! That's because the recipes have been copied to the Flow, but their inputs need to change. Open the copied Prepare recipe "compute_roasteria_training_piiremoved_1" from the Flow view. Then click on the Input/Output tab and change the input dataset.



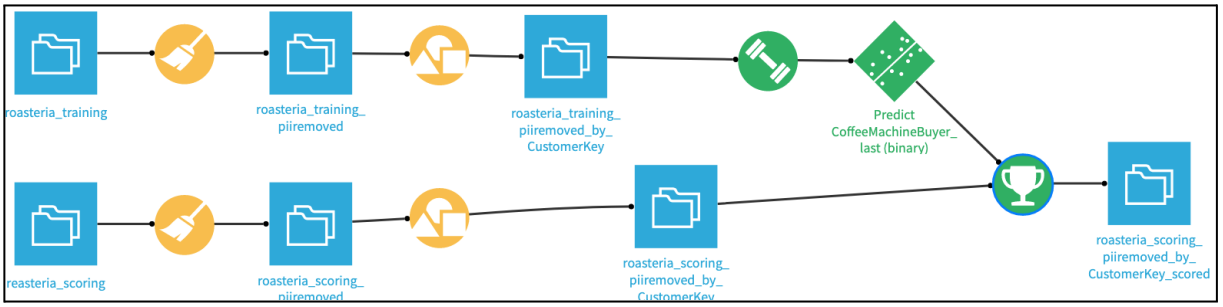
d. Select the "roasteria_scoring" dataset. Save the script. Update the schema, if asked. Then run the script.

e. Run also the second grouping script to produce the final scoring dataset. Select the script from the Flow and click on the run button. When asked, check the "Recursive" option and then "Build Dataset".



f. You now 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_piiremoved_by_..." as the input dataset. Click Create Recipe.

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



h. 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).

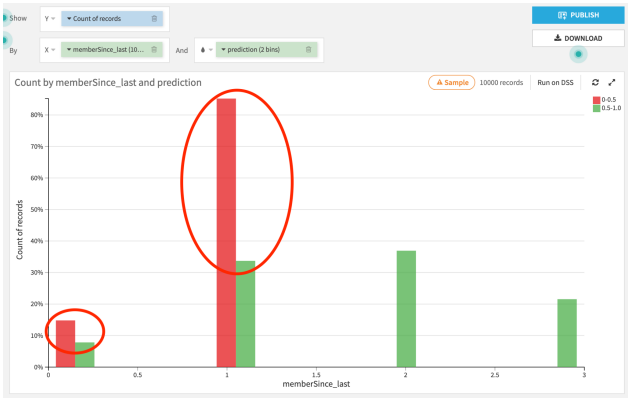
i. Use the Analyze tool on the prediction column to see what percentage of new customers the model expects to buy a coffee machine.

5- Analyze the characteristics of Non-Buyers

a. Now that you have scored new customers it is time to analyze the characteristics of those who do not buy a coffee machine. There are several ways to possibly do this. We advise you to build a few charts to study the typical characteristics of this sub-population of customers.

b. Open the last 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). Onb 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.

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



d. 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.

6- Build the Persona of the Non-Buyers and define your marketing campaign

a. It is now time to wrap up the analysis and define the prototypical profile of your non-buyer customer. Fill the "ANSWER Persona and marketing recommendation EN" document by studying the different profile variables.
b. Once you have identified the characteristics of the profile of the non-buyers 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 non-buyer.
c. Lastly, explain how you would deploy your marketing campaign. How would you make your customer aware of this marketing campaign?

—