

Business Analytics With XLKitLearn: An Excel Frontend For scikit-learn

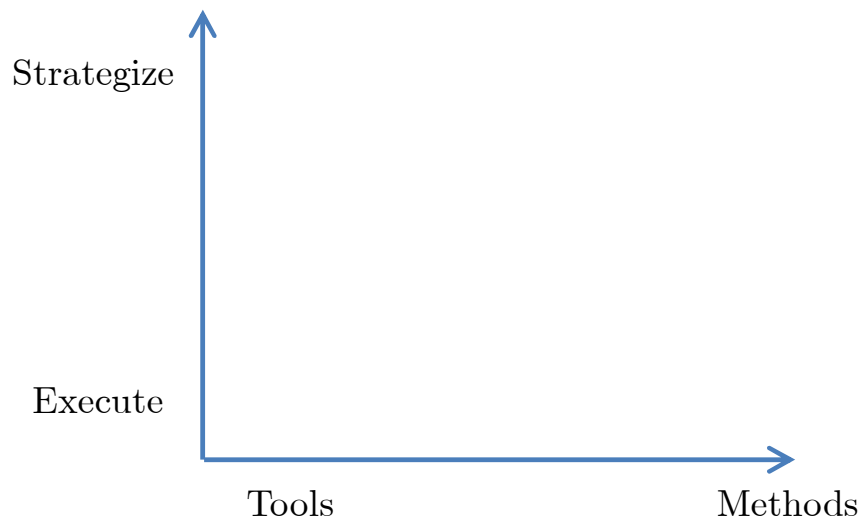
Daniel Guetta
Columbia Business School
guetta@gsb.columbia.edu

May 28th 2021

© 2021 Business Analytics

1

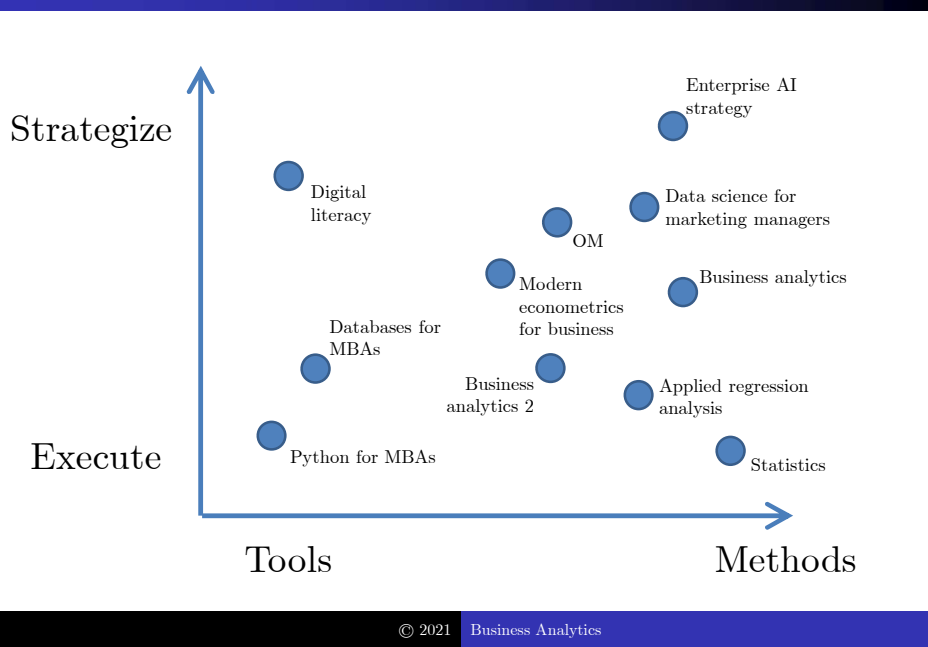
Analytics for MBAs



© 2021 Business Analytics

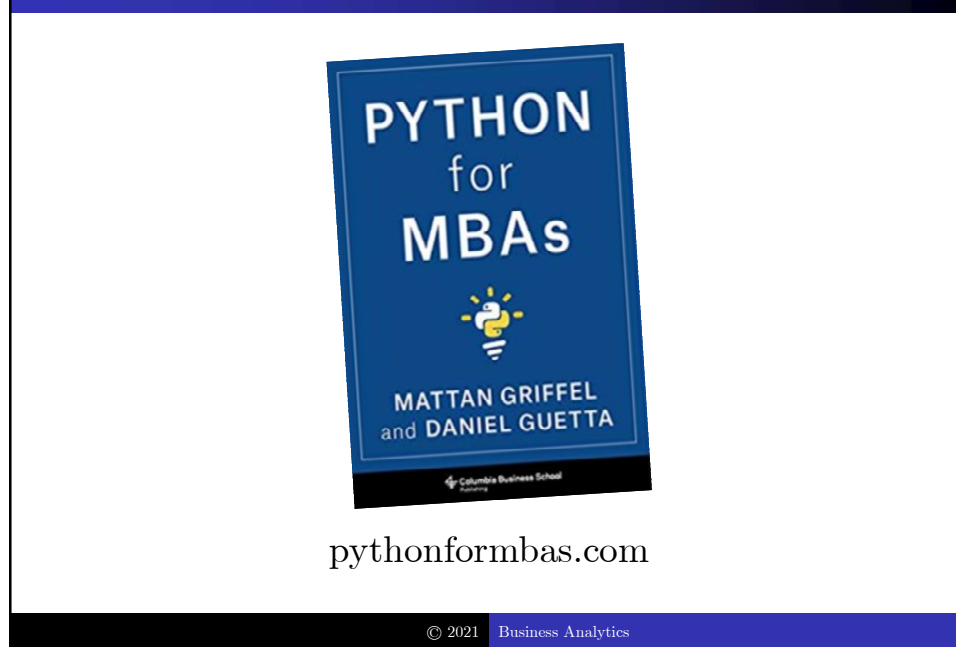
2

Analytics for MBAs

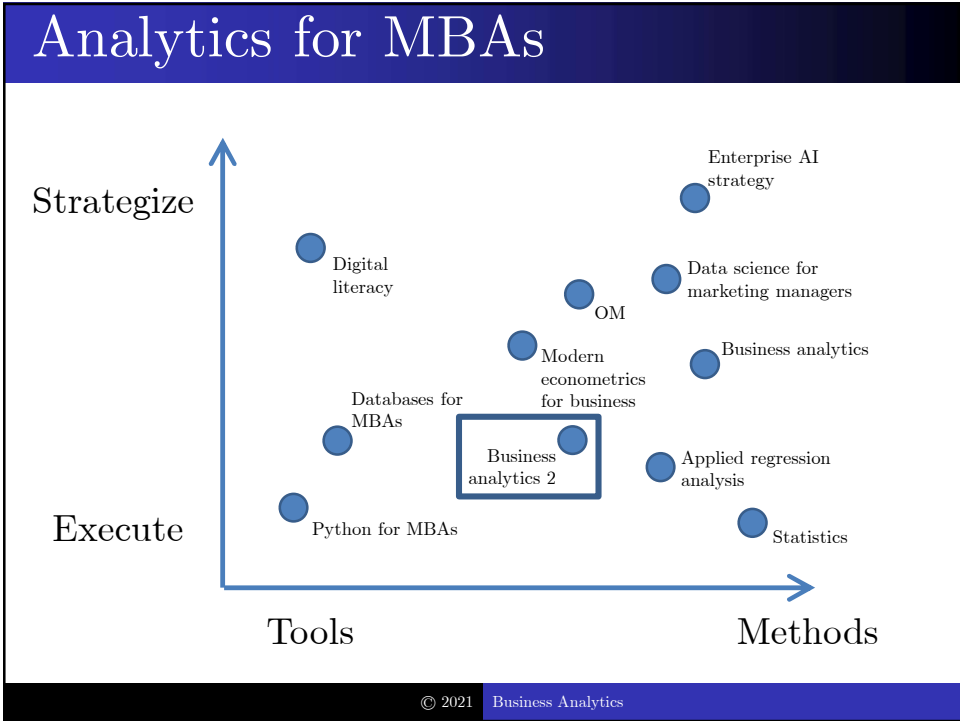


3

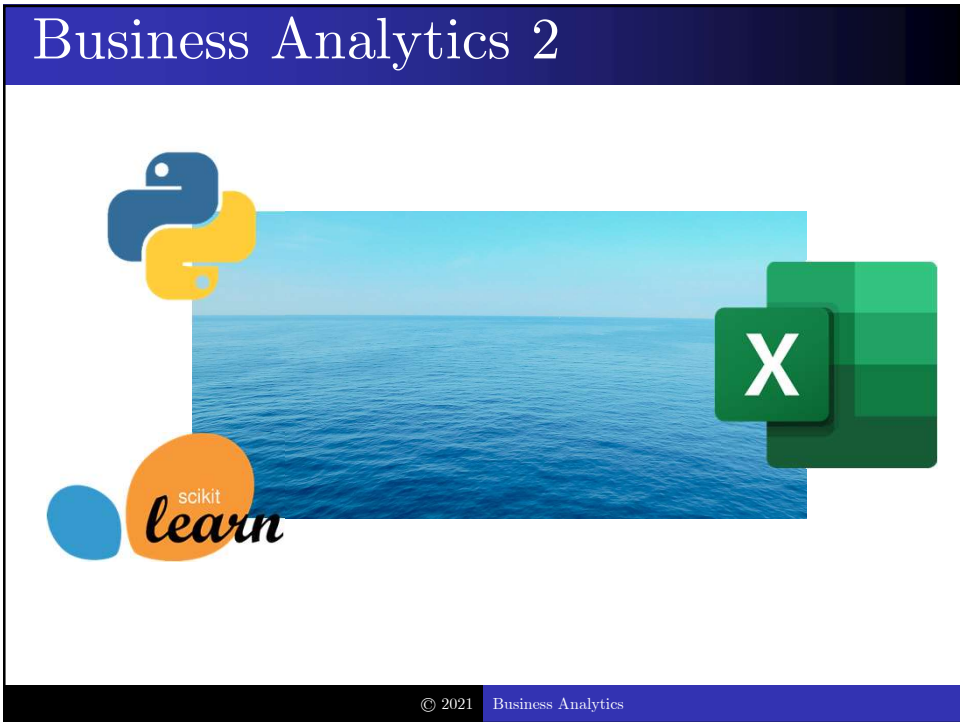
Analytics for MBAs



4



5



6

Business Analytics 2

Potential options

- Use one of many off-the-shelf software products?
 - Black boxes
- Use R/Python and give students code to copy/paste?
 - Poor experience
- Cover everything theoretically without hands-on work?
 - Can be difficult to absorb; dry

© 2021 Business Analytics

7

Our Solution: XLKitLearn



danguetta.github.io/xlkitlearn

© 2021 Business Analytics

8

Predictive analytics

The screenshot shows the 'XLKitLearn Excel Add-in' dialog box. It is titled 'BA2 Excel Add-in' and 'XLKitLearn Excel Add-in'. A note at the top states: 'Any textboxes in red below contain errors. To see the error message, hover over the cell.' The interface is divided into several sections:

- Model Details:** Model is set to 'Random forest'. Formula is 'median_property_value ~ crime_per_c'. The 'Output model' checkbox is checked.
- Parameter(s):** Tree depth is '3 & 4 & 5 & 6'. Number of trees is '25'.
- Training:** Training data is '[xlkitlearn.xlsx]boston_housing!\$A\$1:\$L\$507'. Randomization seed is '123'. The 'Output code' checkbox is checked.
- Testing & Model Selection:** 'Use K-fold cross-validation with 5 folds'.
- Evaluation:** 'Automatically generate an evaluation set with 30 % of the training data' is selected. 'Output evaluation dataset' is checked.
- Prediction:** 'Make predictions for new data' is selected.

A 'Save' button is located at the bottom right of the dialog.

© 2021 Business Analytics

9

Text analytics

The screenshot shows the 'XLKitLearn Text Mining Add-in' dialog box. It is titled 'BA2 Text Mining Add-in' and 'XLKitLearn Text Mining Add-in'. A note at the top states: 'Source Data File (must be in same directory)'. The interface is divided into several sections:

- Source Data File:** A red box indicates an error in the file path.
- Feature Extraction:** Min frequency and Max frequency are empty. Max features is set to '500'. 'Remove English stop words' is checked. 'TF-IDF', 'Include bi-grams', and 'Stem words (CAREFUL)' are unchecked.
- Output:** 'Raw features, to be used with an evaluation set comprising 0 % of the data (sparse)' is selected. 'Results of LDA with 1 topics (with at most 1 iterations)' is also an option.
- Randomization seed:** '123'. The 'Output code' checkbox is checked.

A 'Save' button is located at the bottom right of the dialog.

© 2021 Business Analytics

10

Each run generates Python

```
Equivalent Python code
40
41 # =====
42 # = Import packages =
43 # =====
44
45
46 import pandas as pd
47 import xlwings as xw
48 import sklearn.model_selection as sk_ms
49 import numpy as np
50 import patsy as pt
51 import sklearn.ensemble as sk_e
52 import sklearn.metrics as sk_m
53
54 # =====
55 # = Load the datasets =
56 # =====
57
58 raw_datasets = {}
59
60 raw_datasets["training_data"] = ( xw.Book("XlKitLearn DEBUG.xlsm")
61                                 .sheets("boston_housing")
62                                 .range("SA$1:SL$507")
63                                 .options(pd.DataFrame, index=0, header=1)
64                                 .value )
65
66 # Split the training dataset into a training and evaluation set. We use the
67 # sklearn function. Note that we need to tell the function to shuffle the data.
68 raw_datasets["training_data"], raw_datasets["evaluation_data"] = (
69     sk_ms.train_test_split(raw_datasets["training_data"],
70                           train_size = 1 - 0.3,
71                           test_size = 0.3,
72                           random_state = 123,
73                           shuffle = True) )
74
```