WALMART – DATA HACK

October 1, 2016

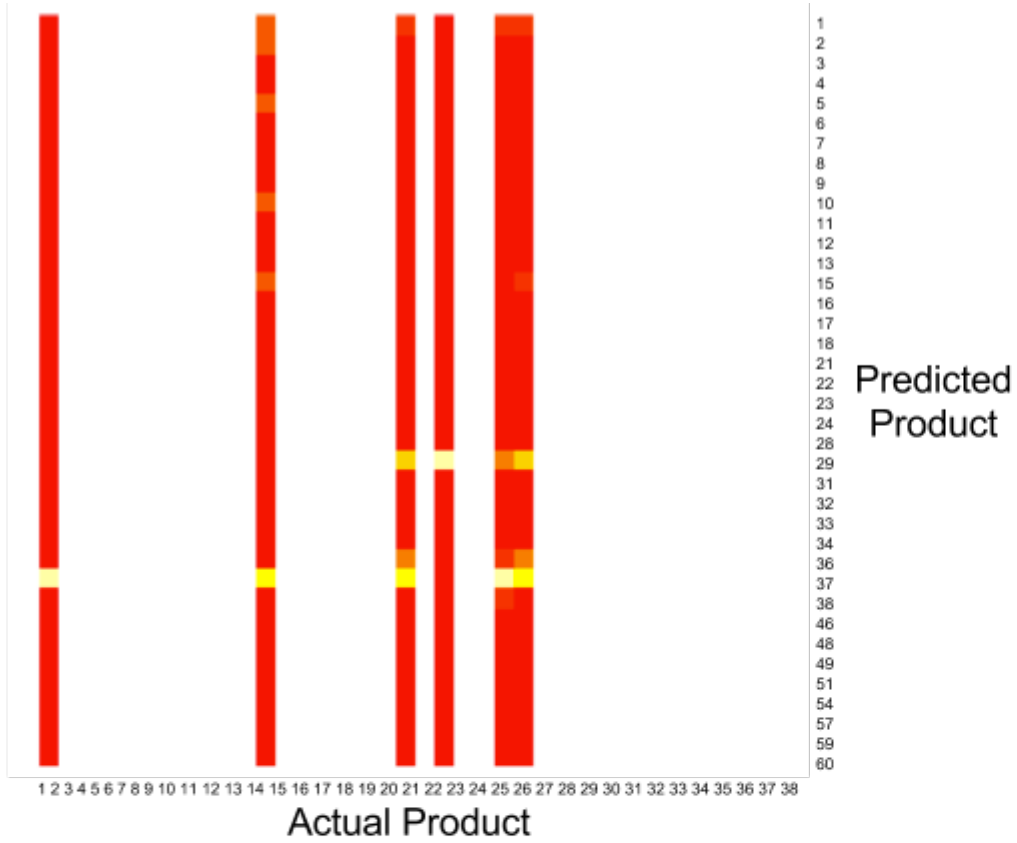
Stephanie Doctor
Rachel Zhang
Amla Srivastava
Sanjmeet Abrol

Business question

What do customers buy when
the most popular product is not available?

Results

- Multiclass logistic regression model that predicts the most popular product (by units sold) from a store's features and the availability of its products
- Overall accuracy: 29.9%



Confusion matrix

Approach

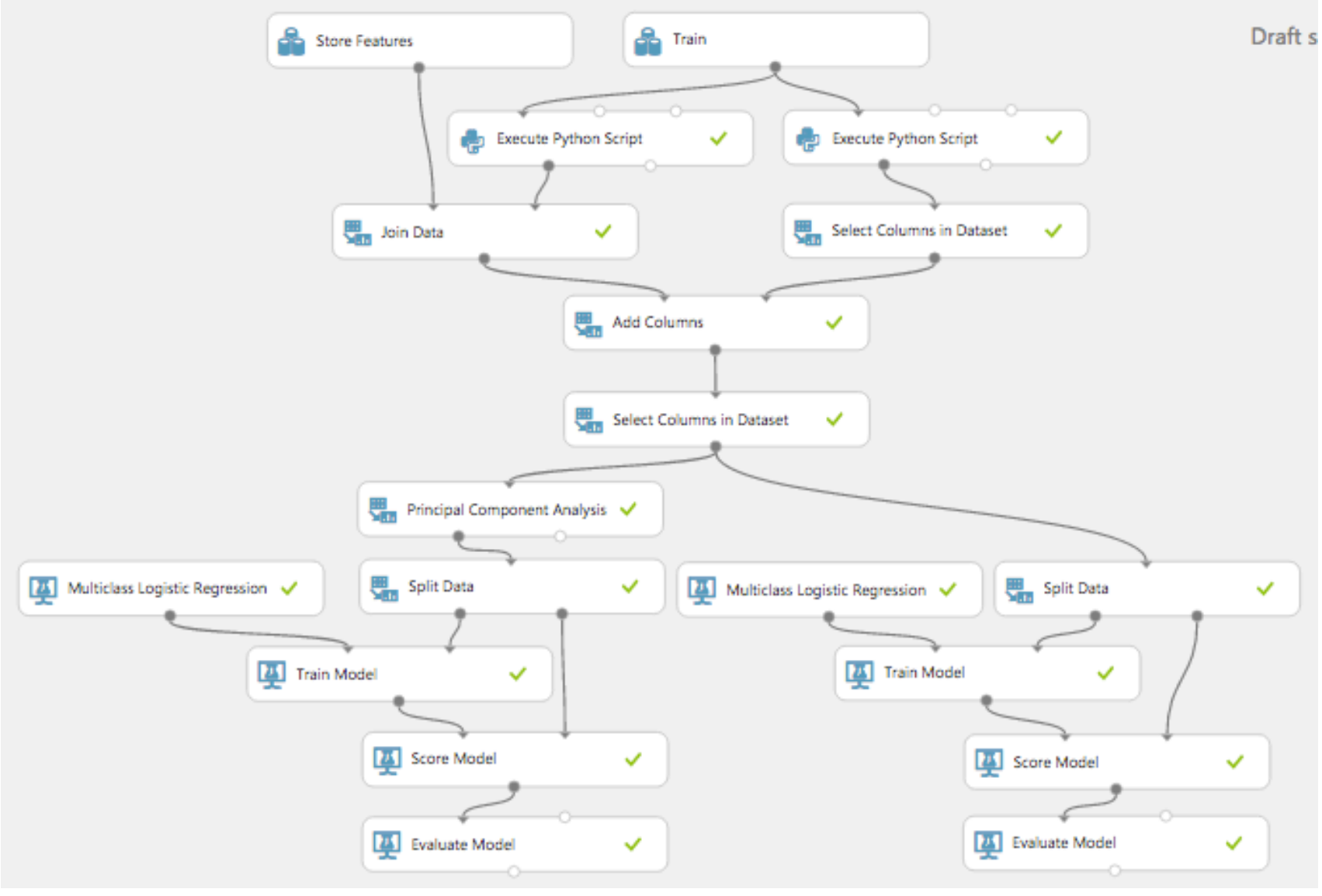
- Determine the most salient product transaction and store attribute features

- Manipulate data structure to facilitate feature engineering, and clean data of missing values

- Use existing variables to create new, more informative features

- Train and evaluate multiclass logistic regression model

Analysis pipeline



Next steps

- Feature engineering
  - Experiment with different feature transformations
- Model generalization
  - Detect and address overfitting/under-fitting in the model
- Uncovering business insight
  - Identify less frequent but significant relationships
  - Understand data better in terms of business context