**Yellowbrick: Steering Scikit-Learn with Visual Transformers**

By Benjamin Bengfort and Rebecca Bilbro

---

**Which features do I use?**

Given labeled data about rooms...
- Which features are most predictive?
- Empty or occupied?

- Redundancy and Parallel Coordinates
  - Use Yellowbrick Redundancy Visualization of Parallel Coordinates to improve feature understanding.

- Rank2D
  - Given labeled data about credit card default...
  - Feature relationship?
  - Correlation and/or
    - Feature importance
  - Use Yellowbrick Rank2D for pairwise feature analysis?

**The API: Scikit-Learn**

Scikit-Learn has so many models, making automated model selection very convenient!

- From classifier.fit(X, y)
- From classification_report(self)
- From classification_report(self, X, y, target_names)
- From classification_report(self, y, y_pred)
- From classification_report(self, X, y, target_names)
- From classification_report(self, y, y_prob)
- From classification_report(self, X, y)

**Which model should I use?**

- Prediction Error and Residuals Plot
  - Visualize the distribution of error to diagnose heteroscedasticity.

- ROC/AUC, Classification Report, Confusion Matrix, and Class Balance
  - ROC/AUC helps us see overall accuracy; classification heatmap helps distinguish Type I, Type II error; and confusion matrix shows error on a per-class basis. What to do with a low-accuracy classifier? Check for imbalance!

**How do I tune my model?**

- Elbow Curves and Silhouette Scores
  - How do you pick an initial value for k in k-means clustering?
  - How do you know whether to increase or decrease k?
  - Is partitive clustering the right choice?

  Higher silhouette scores mean denser, more separate clusters:

**Enter Yellowbrick**

`sklearn-yb.org`

Yellowbrick is a new Python library that:
- Extends the Scikit-Learn API.
- Enhances the model selection process.
- Provides visual tools for feature analysis, diagnostics & steering.

**Frequency Distributions**

- Frequency Distribution of Top 10 Features in a Corpus (Without Stemming)
- Frequency Distribution of Top 10 Features in a Corpus

**t-SNE**

- t-SNE: Uniform Approach of Document Embeddings
- t-SNE: Uniform Embedding of Document Embeddings

---

*Note: The document contains several visualizations and graphs that are not transcribed here.*