SPARK SUMMIT 2017

VISUALIZATION OF ENHANCED SPARK INDUCED NAIVE BAYES CLASSIFIER

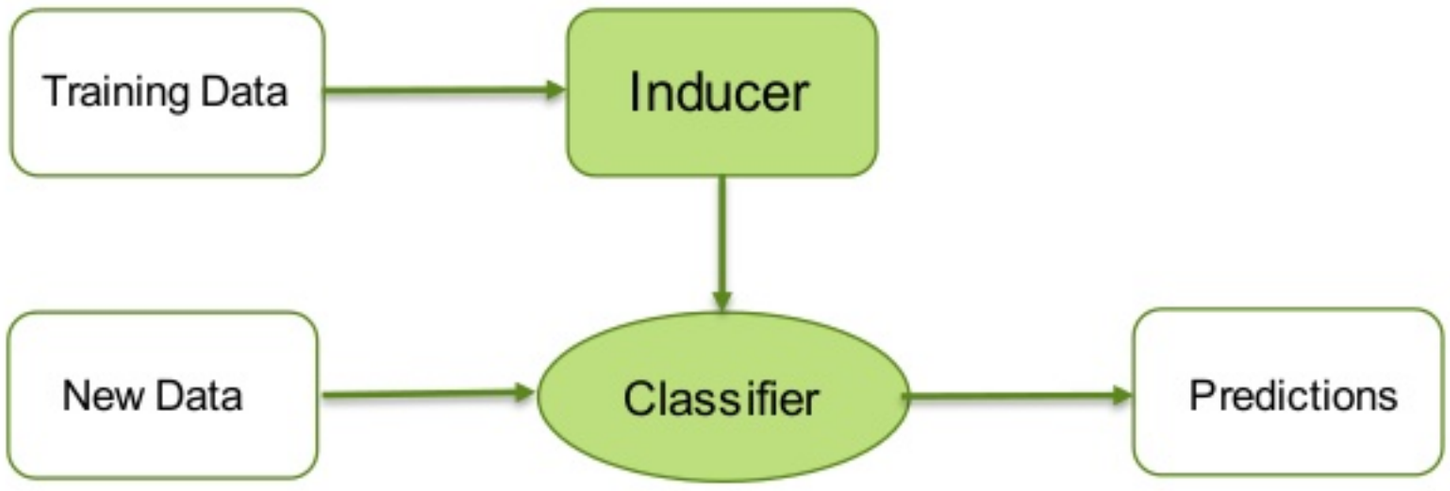
Barry Becker, ESI-Group

Overview

- What the Naïve Bayes Classifier is
- How it can be visualized, and why you would want to
- Limitations of the version in spark, how to get around them
- Demo and use cases

What is a Naïve Bayes Classifier?

- A probabilistic model that can be used to predict a categorical value (the target)
- Induced using training data and evaluated with test data



What is a Naïve Bayes Classifier?

- Advantages
  - Fast and simple
  - Easy to visualize
- Disadvantage
  - Assumes features are independent

Use Case: Detecting Poisonous Mushrooms

Expert determines edibility based on

cap shape, odor, gill spacing, stalk root,
veil type, veil color, ring type, spore print color,
habitat, gill size, stalk shape, … many more….

for 8,124 cases

https://www.pinterest.com/pin/239253798926772284

Bayes' Rule

P(Edible | X) = P(X | Edible) P(Edible) / P(X)

Where X is

Odor = none and

Spore print color = chocolate and

Ring type = evanescent

Bayes Rule Demo

Understand model with Visualization

Everything you see is based on counts

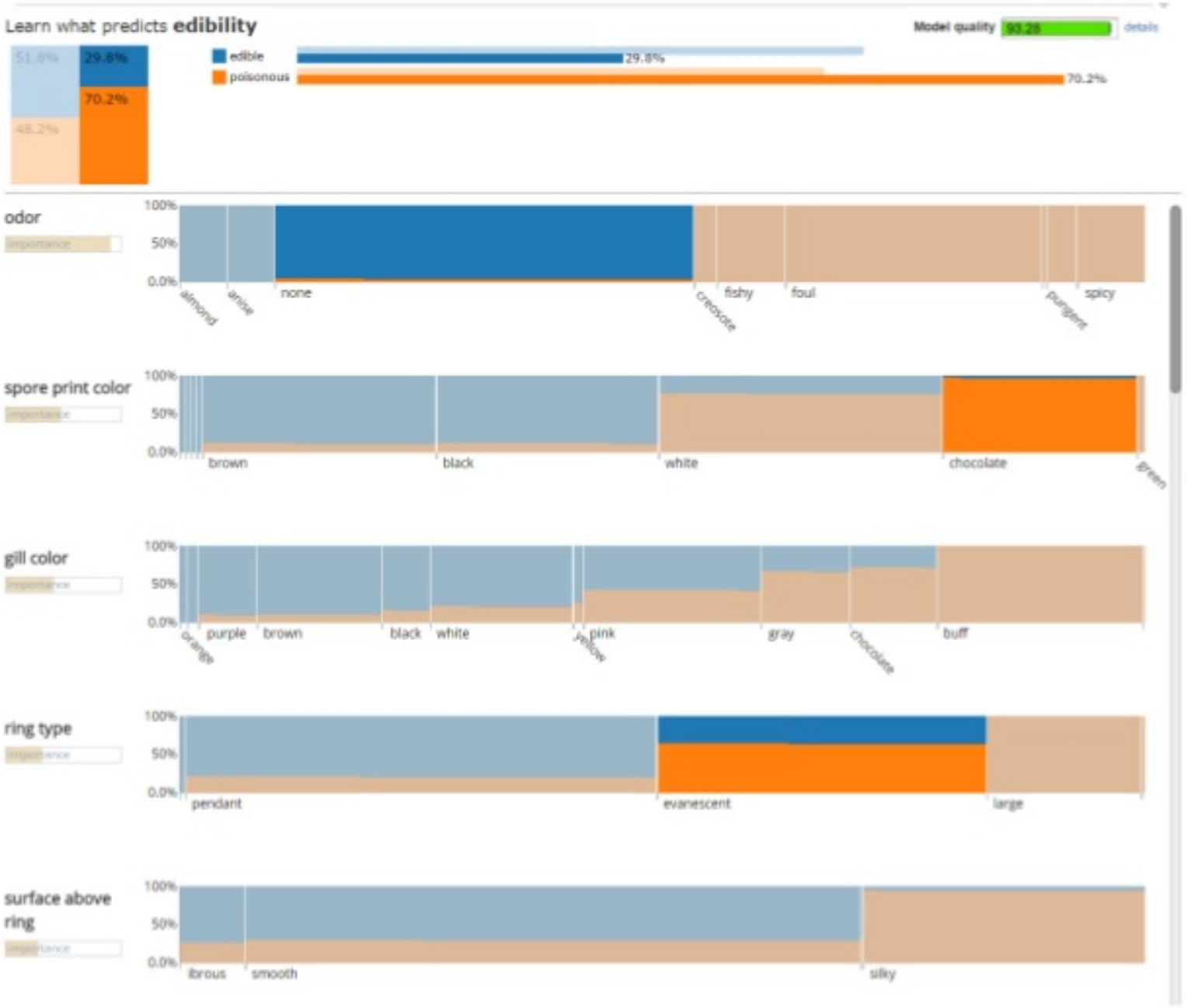
How the Model Makes Predictions

Multiplies conditional probabilities to come up with a posterior probability

SPARK SUMMIT 2017

How the Model Makes Predictions

Multiplies conditional probabilities to come up with a posterior probability

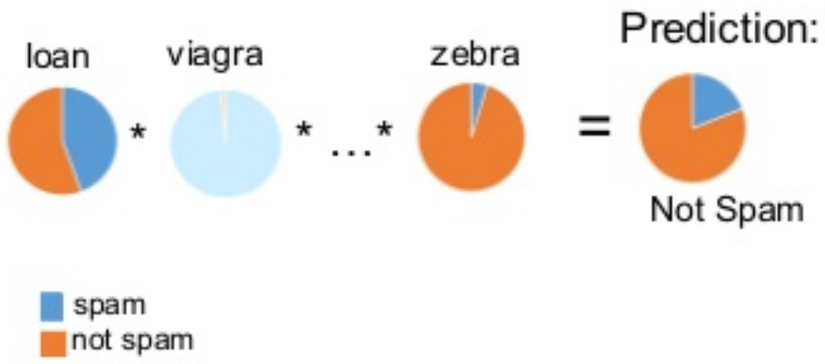


Limitations of Spark Naïve Bayes

Out-of-the-box version is only for document classification!

Document Classification Approach

- Each feature represents a word
- For each feature,
  determine frequency for each class



More General Approach

- Each feature is a column
- For each value of each feature,
  determine class distribution

Limitations of Spark Naïve Bayes

All columns must be non-negative integers

- But what if there are nulls?
- But what if there are strings?
- But what if there are floating point values?
- But what if there are dates?
- What if target is continuous?

How to Handle String Columns?

- First replace Nulls with a special Null value
- Use StringIndexers to map string values to integer indices
- Lastly, use IndexToString transformer to restore the predicted value back to a string.

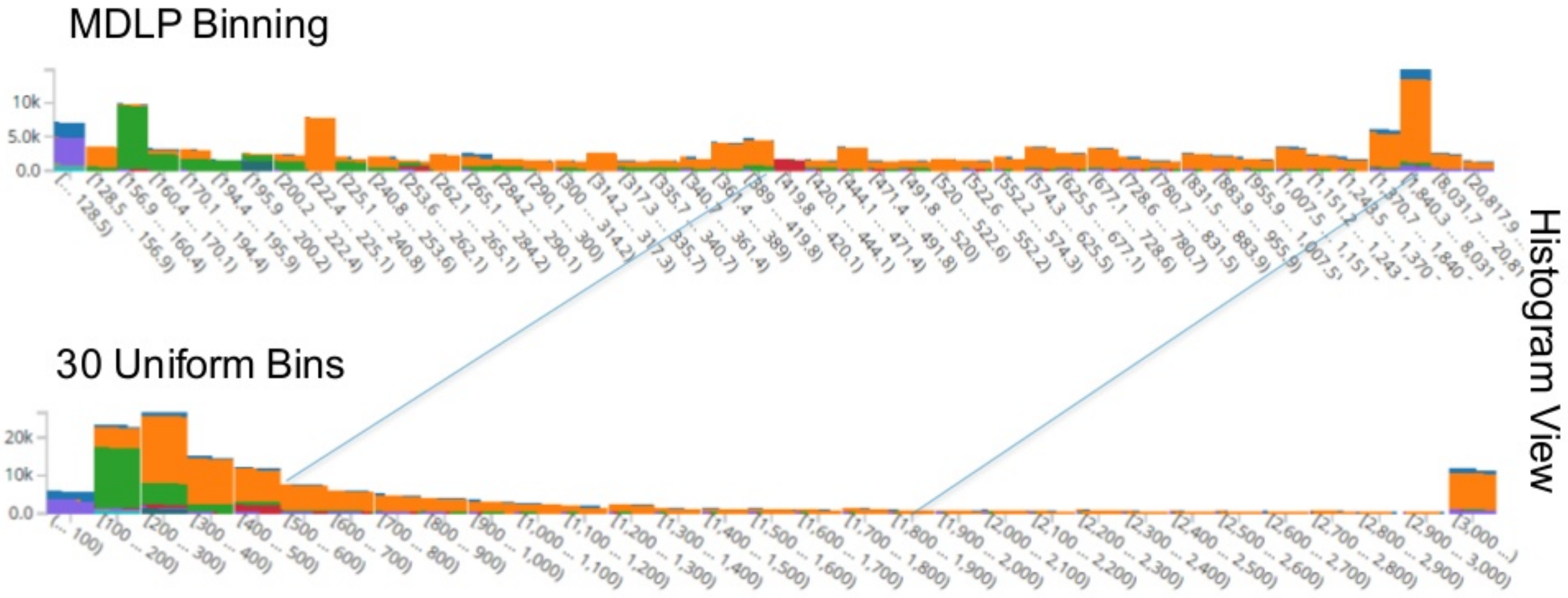
How to Handle Continuous Columns?

- Use MDLP Discretization to intelligently bin continuous (number or date) columns with respect to the target
  - Entropy based binning makes distributions of adjacent bins as different as possible
- Replace Nulls with NaN so they can be in a separate bin.

Example: Job Type from Application

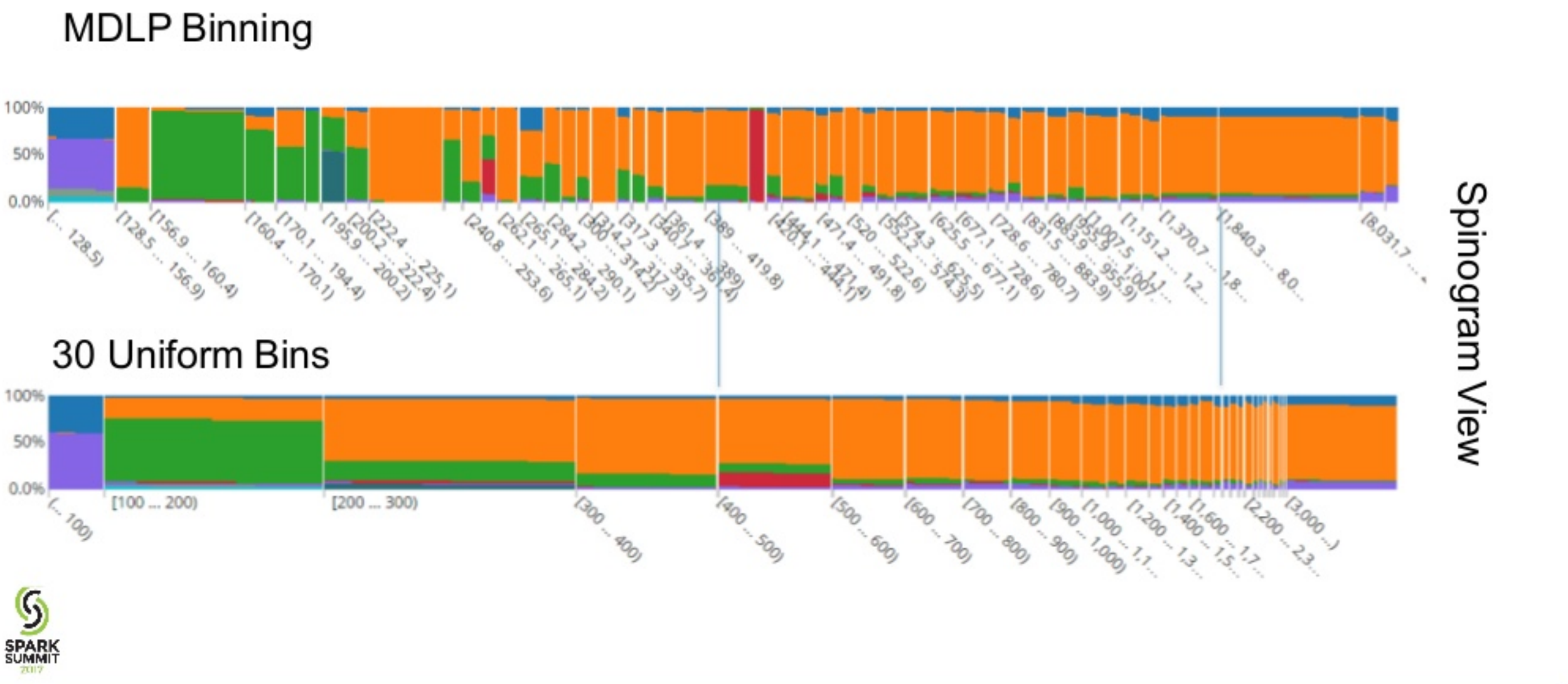
SPARK SUMMIT 2017

MDLP applied to a continuous column



SPARK SUMMIT 2017

MDLP applied to continuous column



SPARK SUMMIT 2017

MDLP Binning of Continuous Features

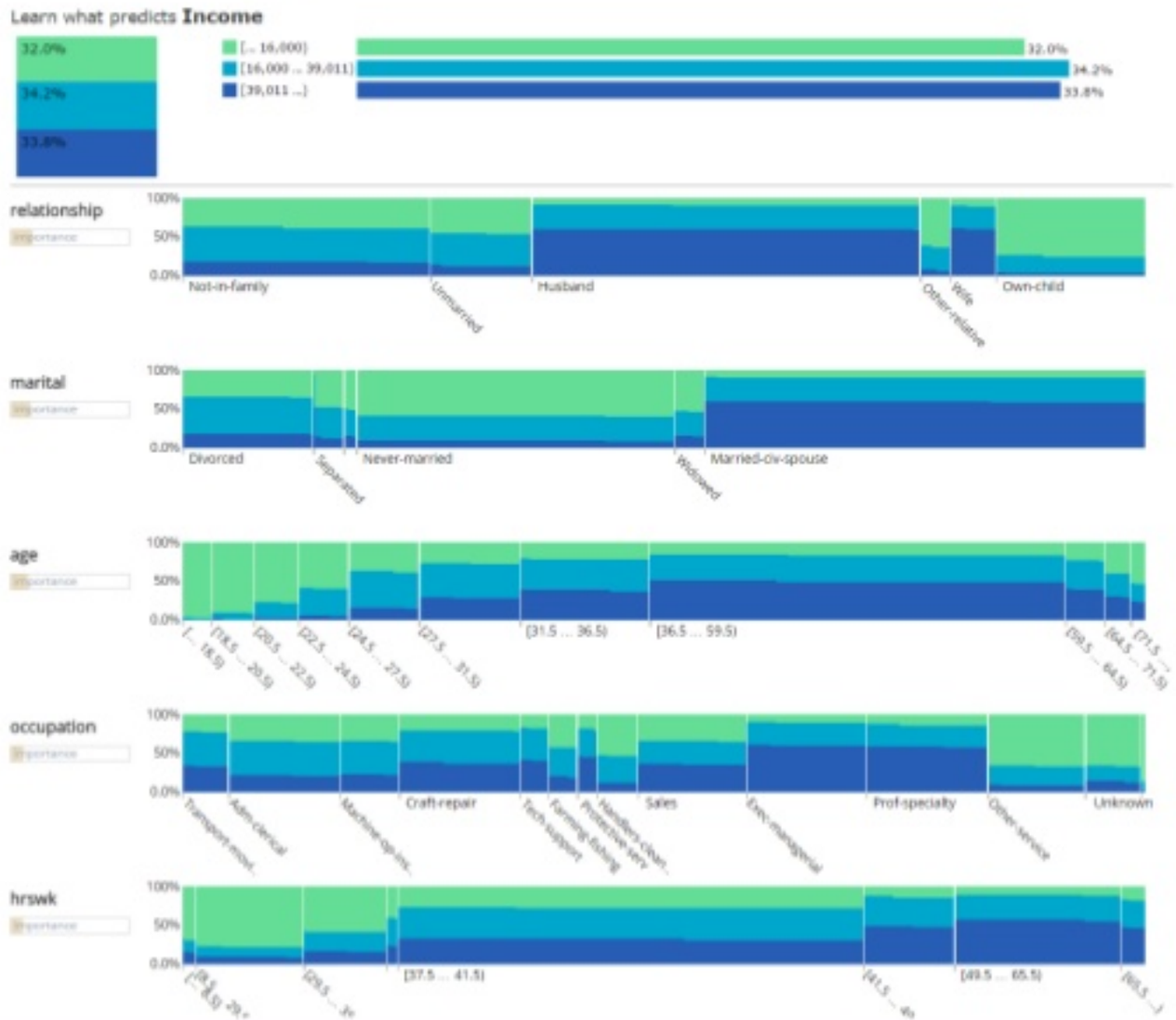
Splits seek to maximize the information, as measured by entropy

Recursively split bins until
- Information gain is too small
- Bins are getting too small (avoids overfitting)

Open Source
https://github.com/sramirez/spark-MDLP-discretization

What if continuous Target?

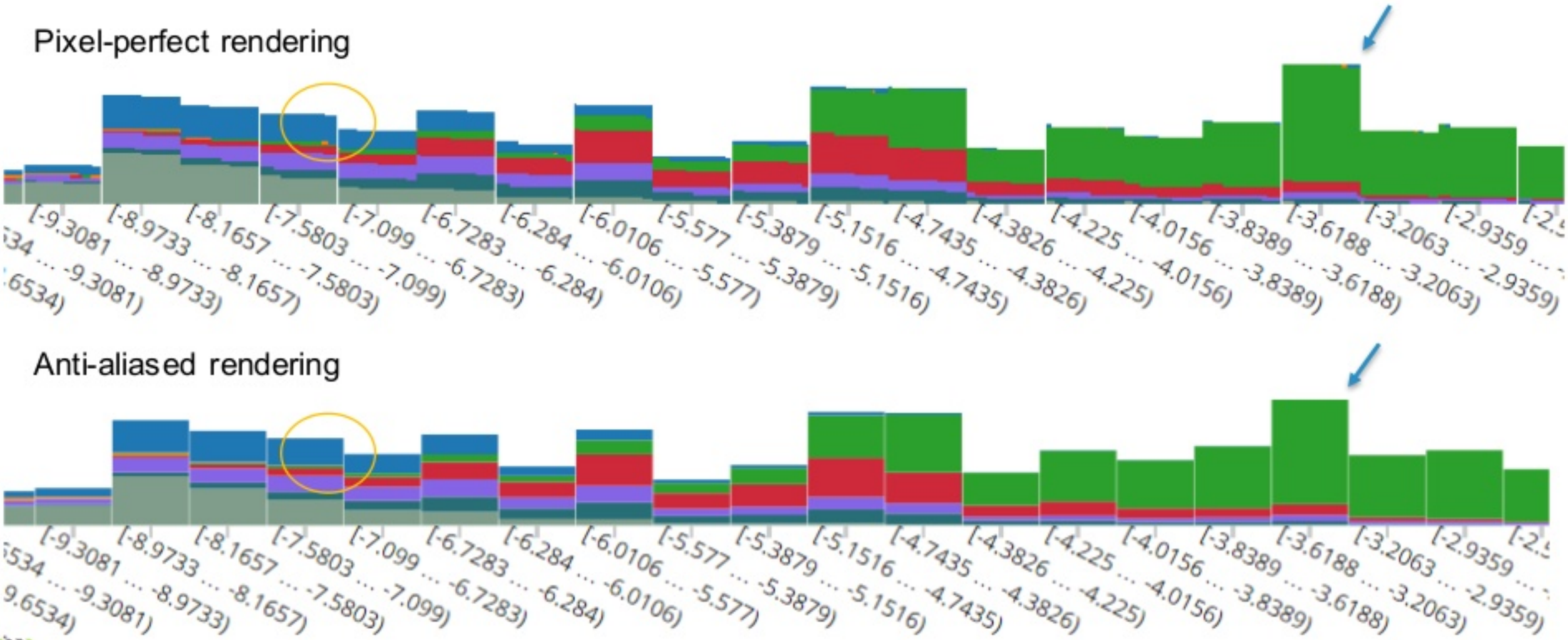
Automatically bin
a continuous target
into 3 equal weight bins



SPARK SUMMIT 2017

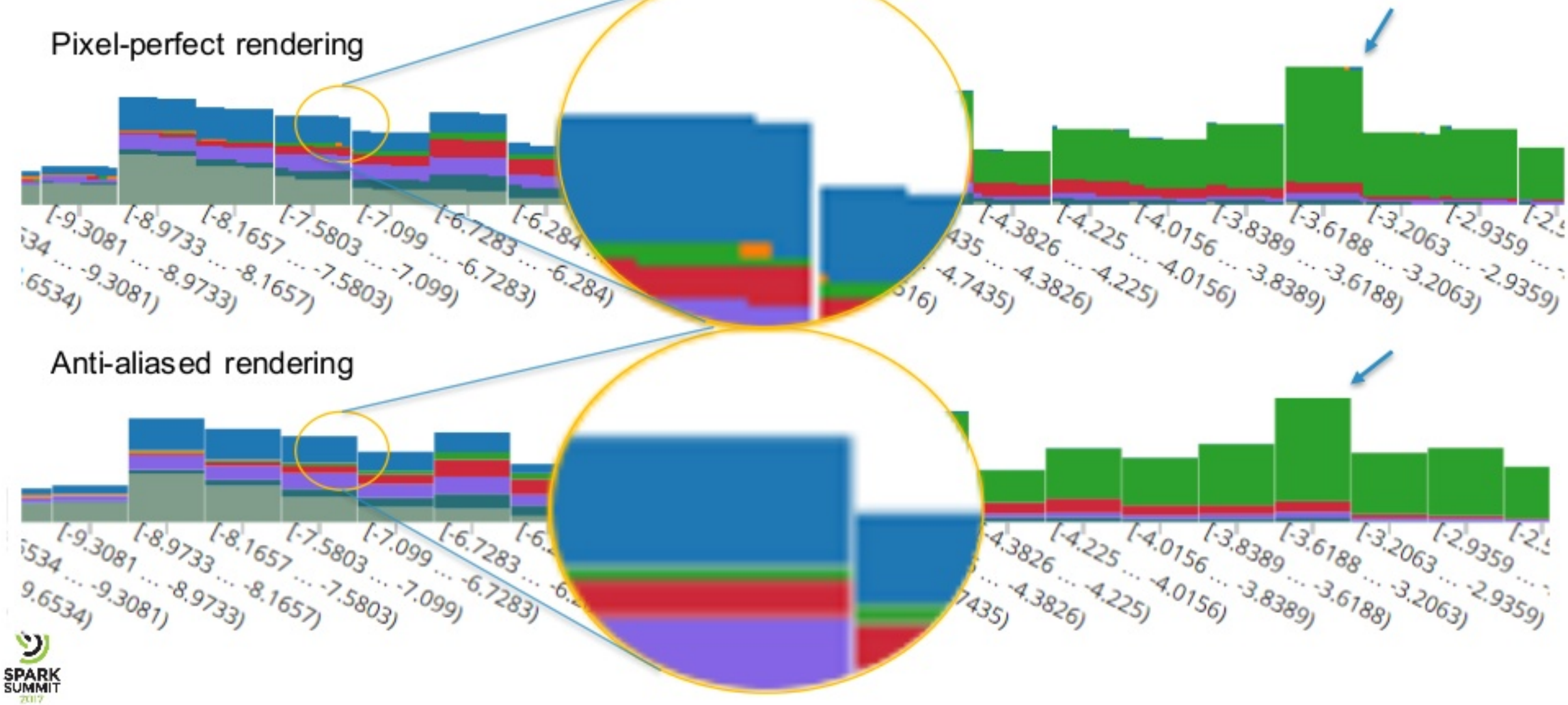
Mineset Demo

Pixel-Perfect Rendering



SPARK SUMMIT 2017

Pixel-Perfect Rendering



SPARK SUMMIT 2017

Conclusion

- Spark Naïve Bayes classifier can be applied to any data as long as some modifications made
- MDLP is useful for binning continuous features
- Visualization of model provides insight, trust, and ability to answer what If questions

https://cloud.esi-group.com/analytics

Barry Becker - bbe@esi-group.com