SESUG
OCTOBER 23-25TH, 2022
MOBILE • ALABAMA

DON'T BE SO ONE-DIMENSIONAL: HOW TO ENGINEER MULTI-DIMENSIONAL HIGH CARDINALITY CATEGORICAL INPUTS FOR MACHINE LEARNING

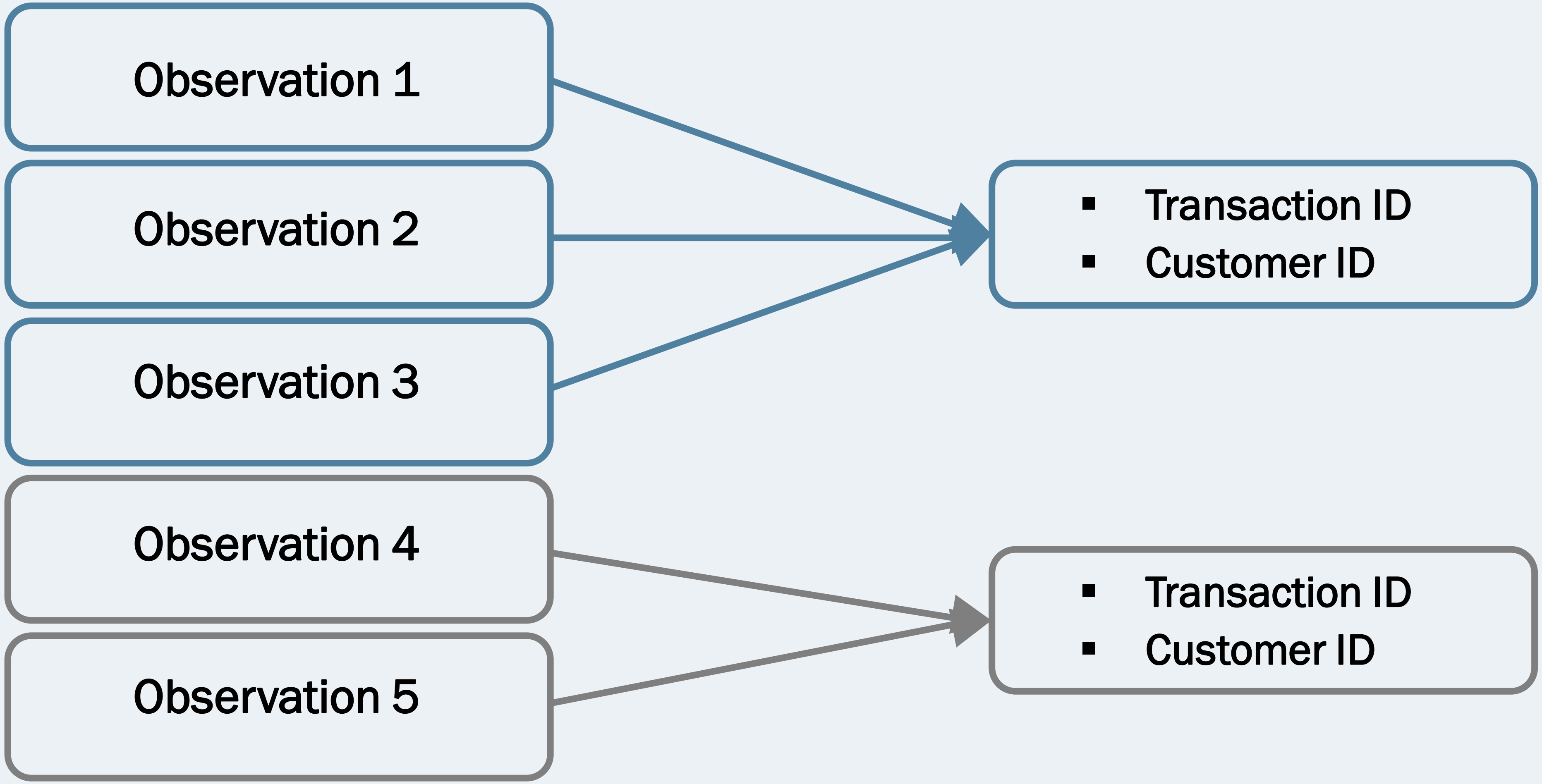
Aleksandar Nikolic, Georgia-Pacific LLC

Outline for Today

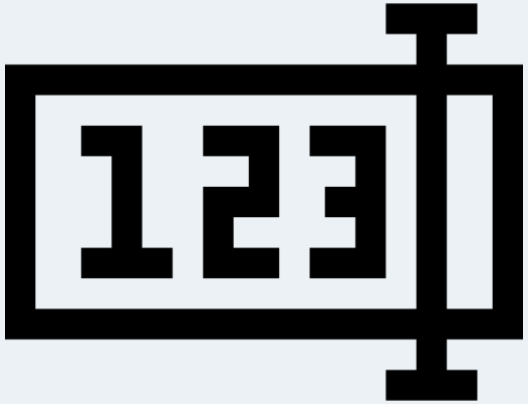
- What is multi-dimensional high cardinality?
- Traditional methods of categorical feature engineering
- New categorical feature engineering method
- K-fold target encoding
- Results of training machine learning models with new features
- Hazards of new method

SESUG
OCTOBER 23-25TH, 2022
MOBILE • ALABAMA

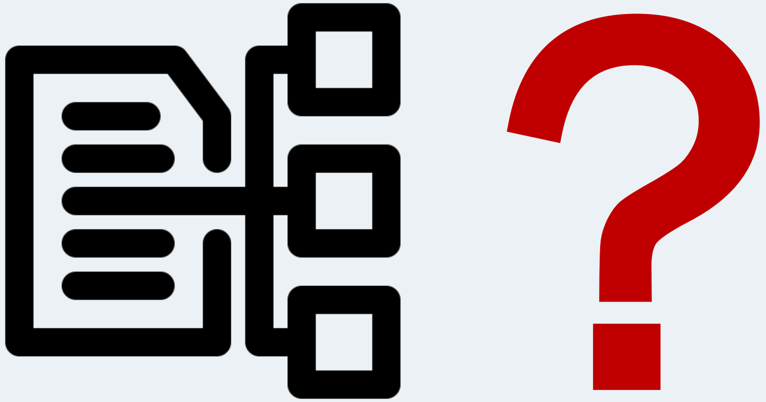
Introduction



There are Many Methods to Easily Summarize Numeric Data, But Not for Categorical Data



- Sum
- Mean
- Min/Max
- Range

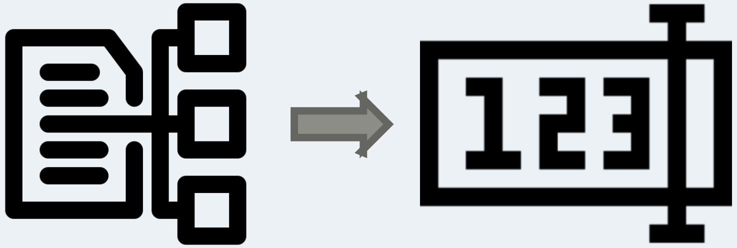


Icon made by Freepik from www.Flaticon.com

Icon made by orvipixel from www.Flaticon.com

SESUG
OCTOBER 23-25TH, 2022
MOBILE • ALABAMA

Aggregating Categorical Data Comes With Multiple Challenges



High cardinality

Inter-level high cardinality + Intra-level high cardinality =

Multi-dimensional high cardinality

What Does a Dataset with Multi-Dimensionality High-Cardinality Look Like?

Inter-level cardinality

Transaction_ID	Product	Fraud
1000	Ruler	1
1000	Binder	1
1200	Desk	0
1200	Stapler	0
1400	Desk	1
1400	Notepad	1
1500	Chair	0
1500	Notepad	0
1500	Envelope	0
2000	Desk	1
2000	Desk	1
2000	Pencil	1
2000	Ruler	1
2000	Pen	1
2000	Highlighter	1

Intra-level cardinality

Traditional Methods of Categorical Feature Engineering Do Not Adequately Solve the Problem of Multi-Dimensional High Cardinality

- Target-agnostic
- Target-based

	Accounts For Inter-level High Cardinality?	Accounts For Intra-level High Cardinality?	No Large Increase of New Inputs?	Conversion to Numeric Input?
Decision Tree Consolidation	✓	✗	✓	✓
One-hot Encoding	✓	✓	✗	✓
String Concatenation	✗	✓	✓	✗

Multi-Dimensional High Cardinality Calls for a Different Approach to Feature Engineering

Transaction_ID	Product	Fraud
1000	Ruler	1
1000	Binder	1
1200	Desk	0
1200	Stapler	0
1400	Desk	1
1400	Notepad	1
1500	Chair	0
1500	Notepad	0
1500	Envelope	0
2000	Desk	1
~~2000~~	~~Desk~~	~~1~~
2000	Pencil	1
2000	Ruler	1
2000	Pen	1
2000	Highlighter	1

- Target encoding
  - For “Desk”, the target encoded value would be .75 (3÷4)
- **Target representation encoding**
  - The target representation value for “Desk” would be .3 (3÷10)
  - Double counting
    - Solved by **de-duplicating observations** at the *Transaction_ID* and *Product* levels
    - New value for “Desk” would be .22 (2÷9)

Why Use Target Representation Instead of Target Encoding?

Transaction_ID	Product	Fraud
1000	Ruler	1
1000	Binder	1
1200	Desk	0
1200	Stapler	0
1400	Desk	1
1400	Notepad	1
1500	Chair	0
1500	Notepad	0
1500	Envelope	0
2000	Desk	1
2000	Pencil	1
2000	Ruler	1
2000	Pen	1
2000	Highlighter	1

- Target representation encoding **places less weight** on **rarely occurring categorical values**
  - For "Pen", the target encoded value is **1** (**1÷1**) while the target representation value is **.11** (**1÷9**)
  - For "Desk", the target encoded value is **.66** (**2÷3**) while the target representation value is **.22** (**2÷9**)
- Target representation is **more interpretable**

Step 1: Concatenate the Transaction ID and Categorical Input Columns

❶ ❷ ❸

Transaction_ID	Product	Fraud	Transaction_ID_Product
1000	Ruler	1	1000_Ruler
1000	Binder	1	1000_Binder
1200	Desk	0	1200_Desk
1200	Stapler	0	1200_Stapler
1400	Desk	1	1400_Desk
1400	Notepad	1	1400_Notepad
1500	Chair	0	1500_Chair
1500	Notepad	0	1500_Notepad
1500	Envelope	0	1500_Envelope
2000	Desk	1	2000_Desk
2000	Desk	1	2000_Desk
2000	Pencil	1	2000_Pencil
2000	Ruler	1	2000_Ruler
2000	Pen	1	2000_Pen
2000	Highlighter	1	2000_Highlighter

Step 2: Remove Duplicate Transaction_ID_Product Observations and Create a Target Hit Indicator Column

❶ ❷

Product	Raw_Product_TH	Transaction_ID_Product	Transaction_ID
Desk	1	1400_Desk	1400
Desk	1	2000_Desk	2000
Ruler	1	1000_Ruler	1000
Ruler	1	2000_Ruler	2000
Pencil	1	2000_Pencil	2000
Pen	1	2000_Pen	2000
Highlighter	1	2000_Highlighter	2000
Notepad	1	1400_Notepad	1400
Binder	1	1000_Binder	1000

- If *Fraud* = 1, then *Raw_Product_TH* = 1

Step 3: Create Target Representation Column

Product	❶ Tot_Raw_Product_TH	❷ Raw_Product_Target_Rep
Desk	2	22%
Ruler	2	22%
Pencil	1	11%
Pen	1	11%
Notepad	1	11%
Highlighter	1	11%
Binder	1	11%
Totals	**9**	**100%**

- Take the **sum of the target hits** by *Product*, and calculate the **proportion of target hits** by *Product* to get the target representation value (*Raw_Product_Target_Rep*)
- "Totals" row included for clarity, **not necessary for this step**

Step 4: Map Raw Target Representation Column Back to Dataset Created in Step 2

❶

Transaction_ID	Product	Raw_Product_TH_Ind	Raw_Product_Target_Rep
1000	Ruler	1	22%
1000	Binder	1	11%
1400	Desk	1	22%
1400	Notepad	1	11%
2000	Desk	1	22%
2000	Ruler	1	22%
2000	Pencil	1	11%
2000	Pen	1	11%
2000	Highlighter	1	11%

- Use the ***Product*** column as the **key**

Step 5: Create New Features Summarized to the Unique Level of Interest

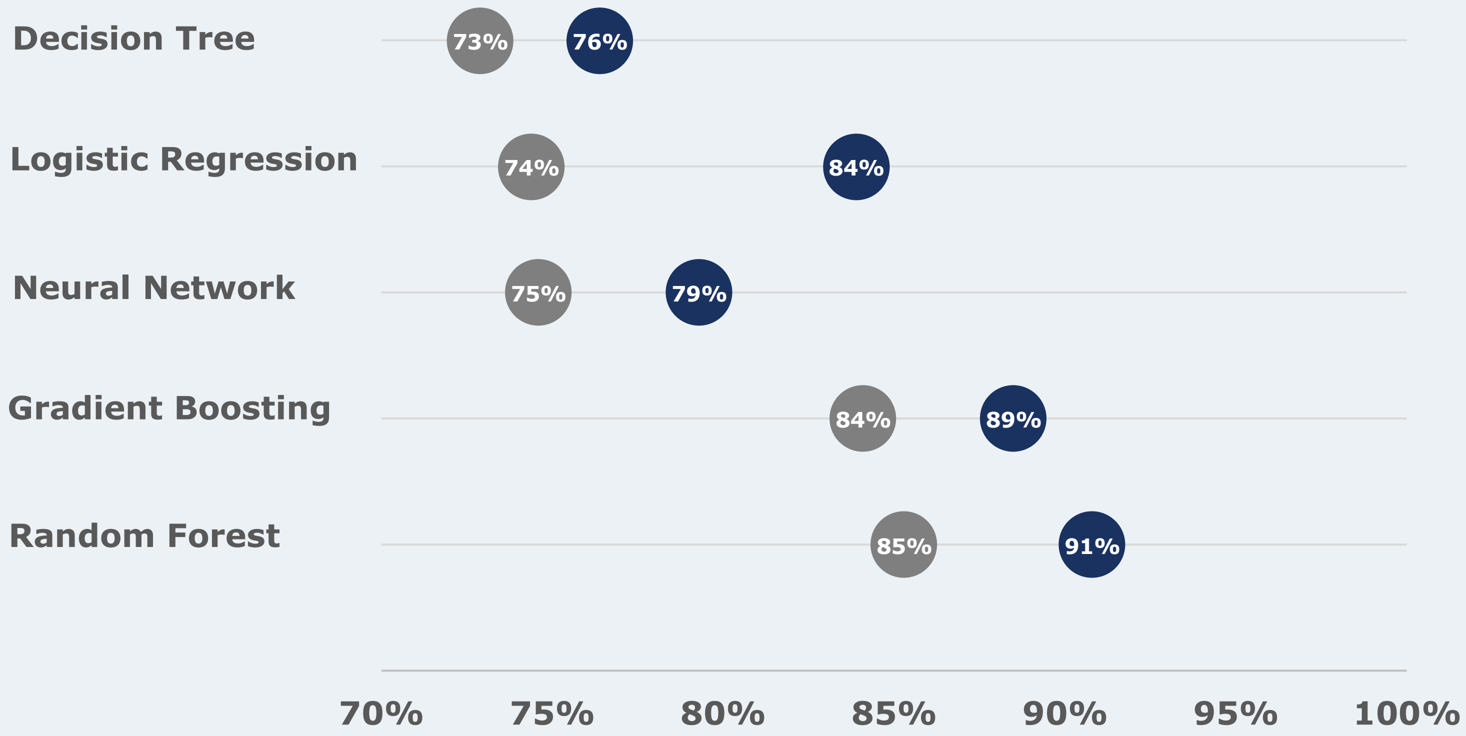
1	2	3	4
Transaction_ID	**Product_Tot_TH**	**Product_Sum_Target_Rep**	**Product_Enc_Prod**
1000	2	33%	0.66
1400	2	33%	0.66
2000	5	78%	3.90

- *Product_Tot_TH* = Sum of the *Raw_Product_TH_Ind* column, count of distinct products associated with target
- *Product_Sum_Target_Rep* = Sum of the *Raw_Product_Target_Rep* column, sum of target representation encodings for all products within level of interest
- *Product_Enc_Prod* = Product of *Product_Tot_TH* and *Product_Sum_Target_Rep* columns

Since the New Features are Based Off the Target Column K-Fold Target Encoding Will Need to be Performed

- Used to head off data leakage and overfitting
- Split the train data into k folds
- Calculate the target representation for each fold
- Calculate the mean target representation from all folds for each unique categorical input value

Including These Newly Created Features Can Improve Machine Learning Model Performance



- **Ten models** were trained, five **with** and five **without** the **four new features**
- All five models **that included** the four features (**blue circles**) **outperformed the models without** the features (grey circles) for recall at the top scored percentile

SESUG
OCTOBER 23-25TH, 2022
MOBILE • ALABAMA

Watchout!

Dirty data

Overfitting

Scoring new data with previously unseen values

Bias towards transactions with multiple observations

Conclusion

- No easy solution for multi-dimensional high cardinality
- Not many solutions publicly available
- New method proven effective on one dataset
- Additional research is necessary
- Always experiment with multiple methods

Thank You For Listening!

Contact Information

Aleksandar Nikolic

Senior Data Scientist

Georgia-Pacific LLC

Email: nikolicxa@gmail.com

The paper, SAS code, and datasets can be found here:

https://github.com/nikolicxa/multi-dimensional-high-cardinality

SESUG
OCTOBER 23-25TH, 2022
MOBILE • ALABAMA