Feature Engineering For Categorical Data

Data Science Applications

Ahmed Basha | Joseph Francis

Supervised by: Felix Neutatz

July 11, 2019

Agenda

CATEGORICAL FEATURES DETECTION

CATEGORICAL FEATURES ENCODING

CONCLUSION

“Automatic Detection For Categorical Features.

(Nominal / Ordinal)”

NOMINAL DATA

Determination of equality. (qualitative)

- ★ *Permutation group*. means any one-to-one substitution.
- ★ *Mode*

ORDINAL DATA

Determination of greater of less.

- ★ *Isotonic group*. means any monotonic increasing function.
- ★ *Median*

NUMERIC DATA

Numerical data is information that is measurable. (quantitative)

- ★ Any mathematical operations can be applied on numerical data.

[*Stevens, Stanley Smith. "On the theory of scales of measurement." (1946): 677-680.*]

Why this mission is not trivial?

Column	Values	Type
City	Berlin, London, Paris	???
Review	Perfect, Good, Bad	???
Student_ID	1, 2, 3, 4 ….	???
Player_Num	12, 8, 23 ….	???
Temperature	36.5, 32, 28.2, 26.6 ….	???

FIRST PHASE WORKFLOW EXAMPLE

Input Dataset		
Name	**Age**	**Size**
John	67	Large
Mark	12	Medium
Chris	22	Small

Labeling Ground Truth	
Columns	**Label**
Name	Nominal
Age	Numerical
Size	Ordinal

Data Imputation	
Categorical	**Numerical**
Most Frequent	Mean Value

Features Generation					
	Dist.	**Freq.**	**W.2vec**	**Binary**	**D.type**
Name	0.25	0.34	86	0	Obj.
Size	0.33	0.77	36	1	Obj.
Age	0.67	0.39	65	0	Int.

Classification	
New Columns	**Predicted Type**
Color	Nominal
Weight	Numerical
Grade	Ordinal

FIRST PHASE DATA & FEATURES

Input Data

11	# of Datasets
127	# of Instances
61	# of Nominal Columns
18	# of Ordinal Columns
48	# of Numeric Columns

Set of Features

Distribution	Is the number of occurrences for each data value.
Frequency	Is the number of unique values to the total number of records.
Word2vec	Vectorizing values, then calculate the distance between values, then take the mean and standard deviation values .
Binary	The attribute has binary values or not.
Data Type	Feature to represent the data type for each attribute.

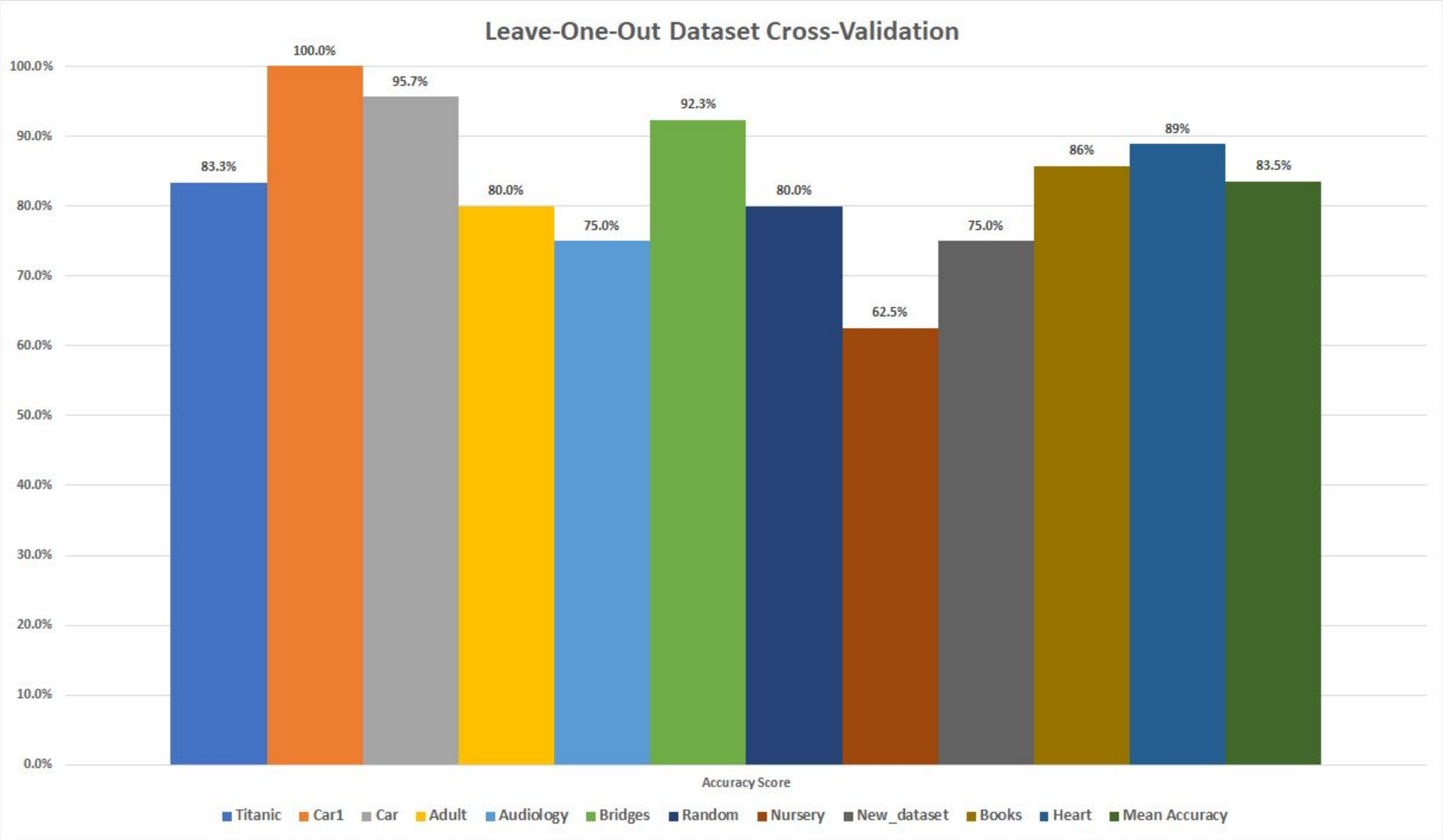
Decision Tree Classifier *Grid Search Hyperparameters*	
• *Criterion*: Gini	• *Max_features*: None
• *Max-depth*: 4	• *Min_samples_split*: 2
• *Splitter*: random	• *Min_samples_leaf*: 2

Leave-One-Out Dataset Cross-Validation

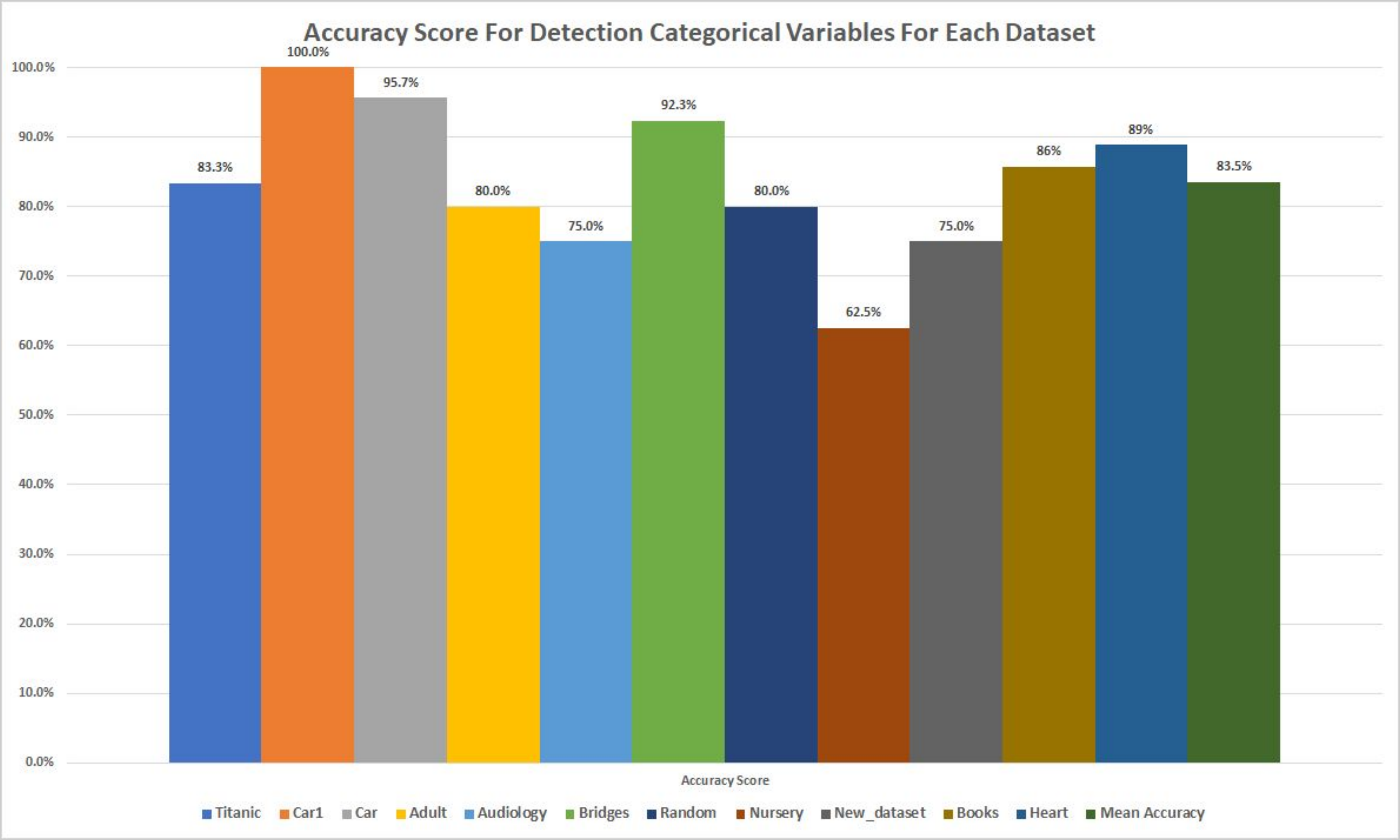
83.5% **Mean Accuracy**

- Training Set: **10 Datasets**
- Testing Set: **1 Dataset**
- Training Model: **Decision Tree Classifier**

FIRST PHASE RESULTS



FIRST PHASE RESULTS



FIRST PHASE EVALUATION

85% *Recall*

Nominal Columns

- Total Count: **61**
- True Positives: **52**
- False Negatives: **19**

50% *Recall*

Ordinal Columns

- Total Count: **18**
- True Positives: **9**
- False Negatives: **9**

100% *Recall*

Numeric Columns

- Total Count: **48**
- True Positives: **48**
- False Negatives: **0**

FIRST PHASE EVALUATION

Examples For False Negatives Predictions				
True Labels	**Dataset**	**Column**	**Predicted Labels**	**Unique Values**
Nominal	Titanic	PassengerId	Numerical	1, 2, 3, 4, …. 890, 891
	Adult	Race	Ordinal	Black, White, Indian ….
	Nursery	Form	Ordinal	Complete, Incomplete, Foster ….
Ordinal	Titanic	Pclass	Nominal	1, 2, 3
	Car	Price	Nominal	Low, Medium, High
	Nursery	Housing	Nominal	Convenient, Less_conv, Critical
	Adult	Education	Nominal	11th,HS-grad, Some-College, 10th, Prof-School, ….

SECOND PHASE TARGET

“Find The Best Feature Representation For Categorical Data That Yield High Model Accuracy.”

SECOND PHASE EXPERIMENTS

1. Encoding Columns Separately Against Target.
2. Encoding All Categorical Columns Together By One Encoder.
3. Encoding One Column With Another Encoder For All Other Columns.
4. Encoding Nominal And Ordinal Columns By Different Encoders.

SECOND PHASE INPUT DATA

Dataset	# of Columns	# of Cat. Columns	More Information
Titanic	12	8	• Unique Values Range: **[2 : 891]** • # of Instances: **891**
Car	23	10	• Unique Values Range: **[2 : 8]** • # of Instances: **203**
Car1	7	5	• Unique Values Range: **[3 : 4]** • # of Instances: **1927**
Bridges	13	9	• Unique Values Range: **[2 : 106]** • # of Instances: **107**
Adult	15	10	• Unique Values Range: **[2 : 41]** • # of Instances: **45222**

SECOND PHASE EXPERIMENTS

1	What will happen if we encode each feature alone without even consider any other features in the model?
2	What will happen if we encode all categorical features together by the exact same encoder?
3	What will happen if we encode only one feature by one encoder while encoding all other features by another encoder at the same time?
4	What will happen if we encode all nominal features by one encoder while encoding all ordinal features by another encoder at the same time?

★ **Encoding Categorical Columns Separately Against Target.**

Brand	Country	Review	Cylinders	Price
BMW	Germany	Perfect	10	72.000$
KIA	Korea	Bad	8	38.000$
Ford	USA	Good	6	57.000$

Categorical Column (Brand)

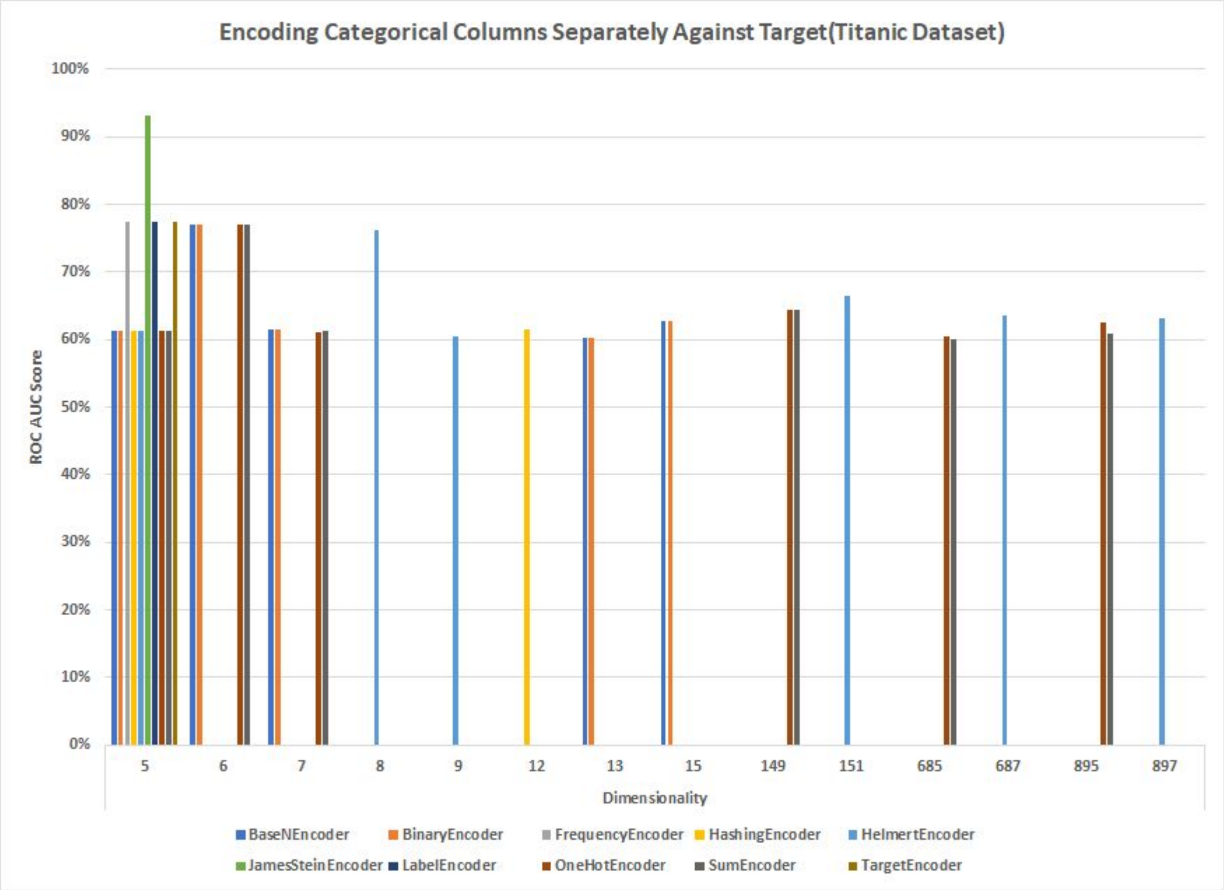
Numeric Column (Cylinders)

Target (Price)

Encoders List

- ★ **Label Encoder**
- ★ Helmert Encoder
- ★ JamesStein Encoder
- ★ Hashing Encoder
- ★ Frequency Encoder
- ★ Binary Encoder
- ★ Target Encoder
- ★ OneHot Encoder
- ★ Sum Encoder
- ★ BaseN Encoder

What happens if we encode each feature alone without even consider any other features in the model?



- ★ JamesStein encoder achieved high score & low dimensionality.
- ★ Helmert encoder produced the highest dimensionality.
- ★ Most encoders produced low dimensionality.

SECOND PHASE EXPERIMENT (2)

★ Encoding All Categorical Columns Together By One Encoder.

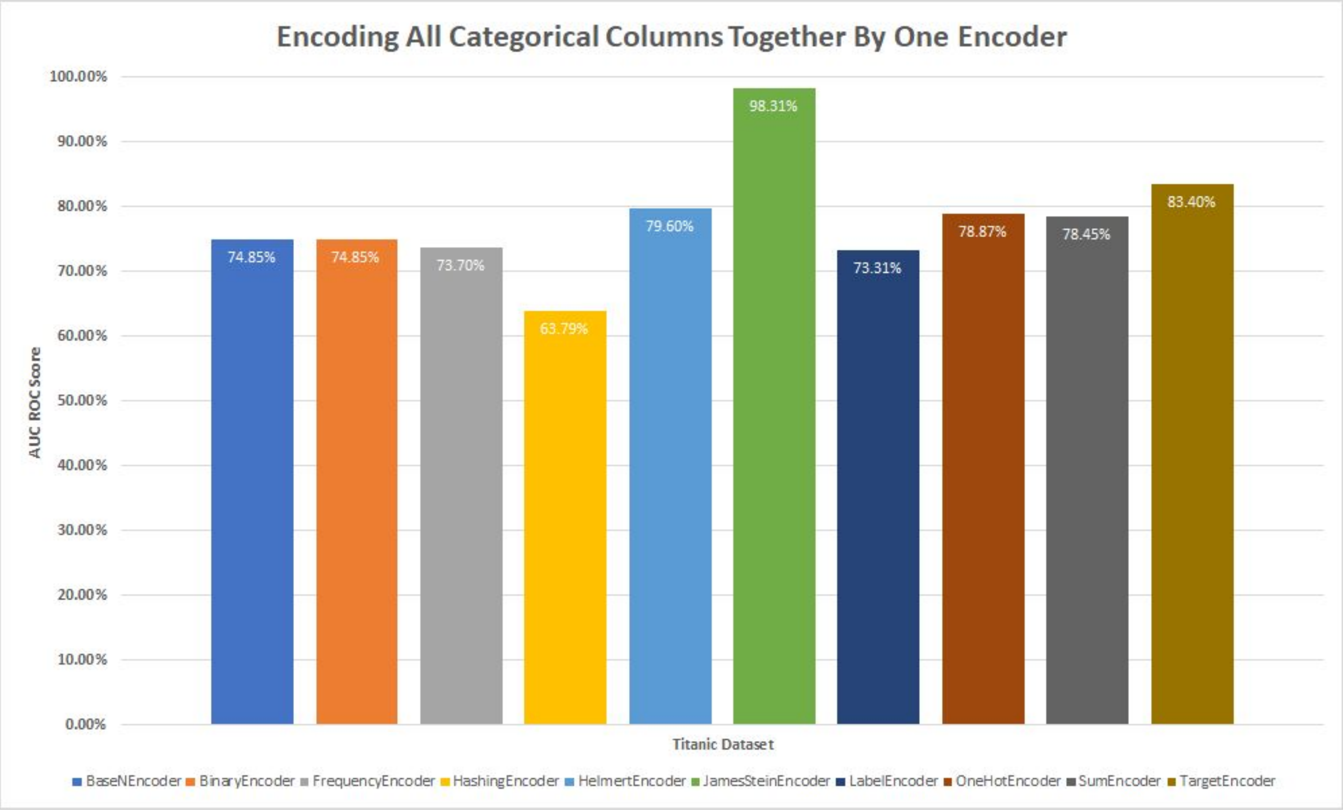
Brand	Country	Review	Cylinders	Price
BMW	Germany	Perfect	10	72.000$
KIA	Korea	Bad	8	38.000$
Ford	USA	Good	6	57.000$

Categorical Columns (Brand, Country, Review) — *Numeric Column* (Cylinders) — ***Target*** (Price)

Encoders List

- ★ **Label Encoder**
- ★ Helmert Encoder
- ★ JamesStein Encoder
- ★ Hashing Encoder
- ★ Frequency Encoder
- ★ Binary Encoder
- ★ Target Encoder
- ★ OneHot Encoder
- ★ Sum Encoder
- ★ BaseN Encoder

What happens if we encode all categorical features together by the exact same encoder?



- ★ Hashing encoder performed extremely bad while JamesStein encoder performed extremely well.
- ★ The rest encoders have approximately the same performance.

SECOND PHASE EXPERIMENT (3)

★ Encoding One Column With Another Encoder For All Other Columns.

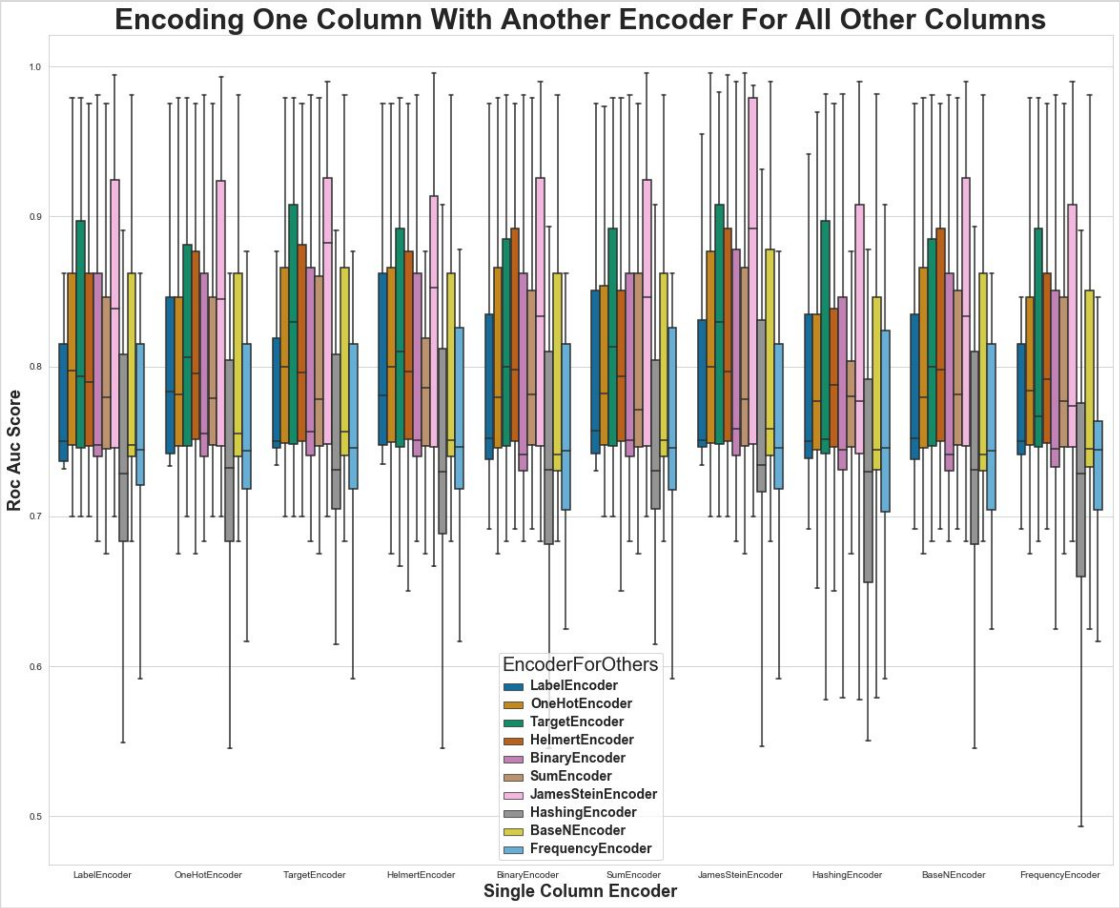
Brand	Country	Review	Cylinders	Price
BMW	Germany	Perfect	10	72.000$
KIA	Korea	Bad	8	38.000$
Ford	USA	Good	6	57.000$

Categorical Column *Other Columns* *Numeric Column* ***Target***

Encoders List

- ★ **Label Encoder**
- ★ Helmert Encoder
- ★ JamesStein Encoder
- ★ Hashing Encoder
- ★ Frequency Encoder
- ★ Binary Encoder
- ★ **Target Encoder**
- ★ OneHot Encoder
- ★ Sum Encoder
- ★ BaseN Encoder

What happens if we encode only one feature by one encoder while encoding all other features by another encoder?



★ Hashing encoder performs extremely bad with nominal columns.

★ JamesStein encoder performs extremely well with nominal columns.

★ Binary and Frequency encoders seems to be not applicable for nominal columns.

SECOND PHASE EXPERIMENT (4)

★ Encoding Nominal And Ordinal Columns By Different Encoders.

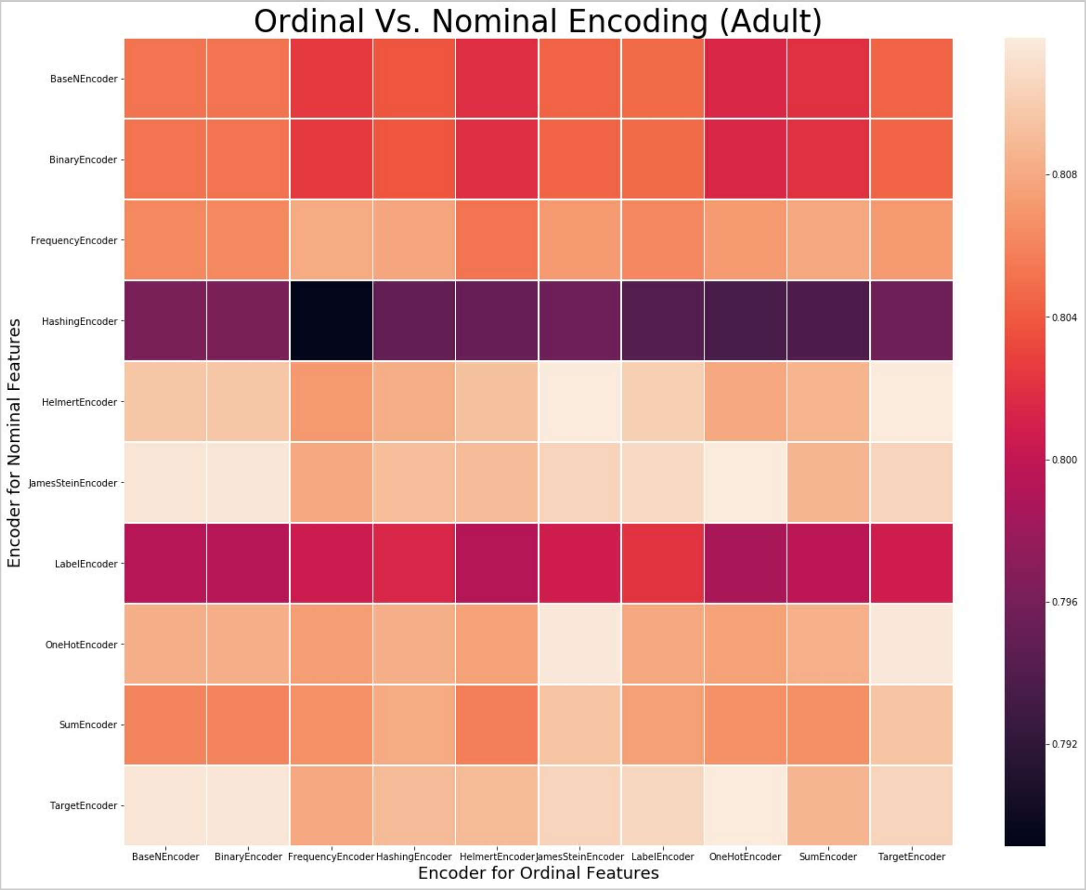
Brand	Country	Review	Cylinders	Price
BMW	Germany	Perfect	10	72.000$
KIA	Korea	Bad	8	38.000$
Ford	USA	Good	6	57.000$

Nominal Columns *Ordinal Columns* *Numeric Column* ***Target***

Encoders List

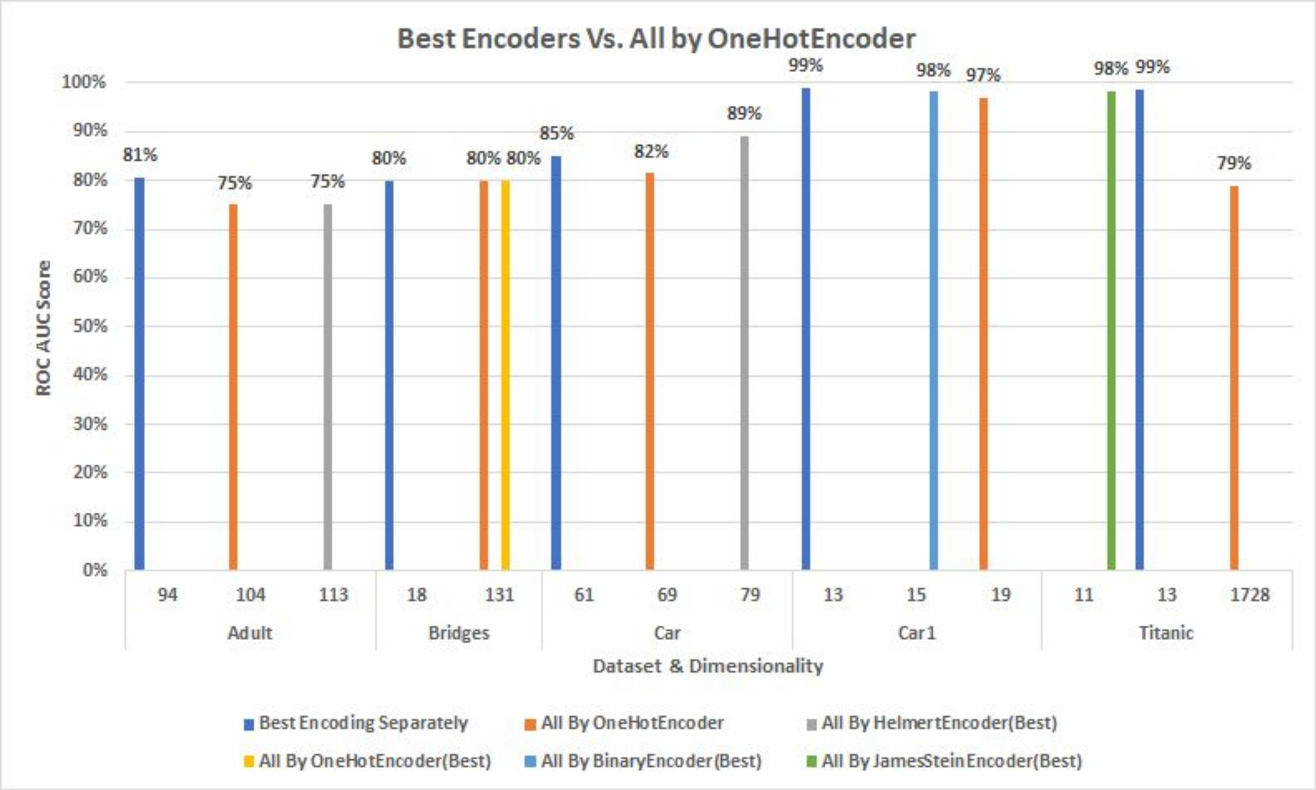
- ★ **Label Encoder**
- ★ Helmert Encoder
- ★ JamesStein Encoder
- ★ Hashing Encoder
- ★ Frequency Encoder
- ★ Binary Encoder
- ★ **Target Encoder**
- ★ OneHot Encoder
- ★ Sum Encoder
- ★ BaseN Encoder

What happens if we encode all nominal features by one encoder while encoding all ordinal features by another encoder?



★ Choosing a feature encoder does not depend on the feature type whether it's nominal or ordinal.

CONCLUSION



★ Carefully choosing the most suitable encoder for each feature leads to low dimensionality and high accuracy.

References

- Katz, Gilad, Eui Chul Richard Shin, and Dawn Song. “Explorekit: Automatic feature generation and selection.” 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.
- Kaul, Ambika, Saket Maheshwary, and Vikram Pudi. “Autolearn—Automated feature generation and selection.” 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017.
- Stevens, Stanley Smith. "On the theory of scales of measurement." (1946): 677-680.
- O'Reilly — Introduction to Machine Learning with Python by Sarah Guido, Andreas C. Müller — Chapter 4. Representing Data and Engineering Features
- Categorical Features and Encoding in Decision Trees — medium.com
- Potdar, K., Pardawala, T.S. and Pai, C.D., 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 175(4), pp.7-9.
- Beyond One-Hot: an exploration of categorical variables. kdnuggets.com

CONCLUSION (Backup) FINDINGS

Highly Recommended

- ★ JamesStein Encoder
- ★ Target Encoder

Recommended

- ★ OneHot Encoder
- ★ Helmert Encoder
- ★ Binary Encoder

Neutral

- ★ Sum Encoder
- ★ Label Encoder
- ★ BaseN Encoder

Not Recommended

- ★ Hashing Encoder
- ★ Frequency Encoder

The Best Feature Representation For Categorical Data That Yield High Model Accuracy

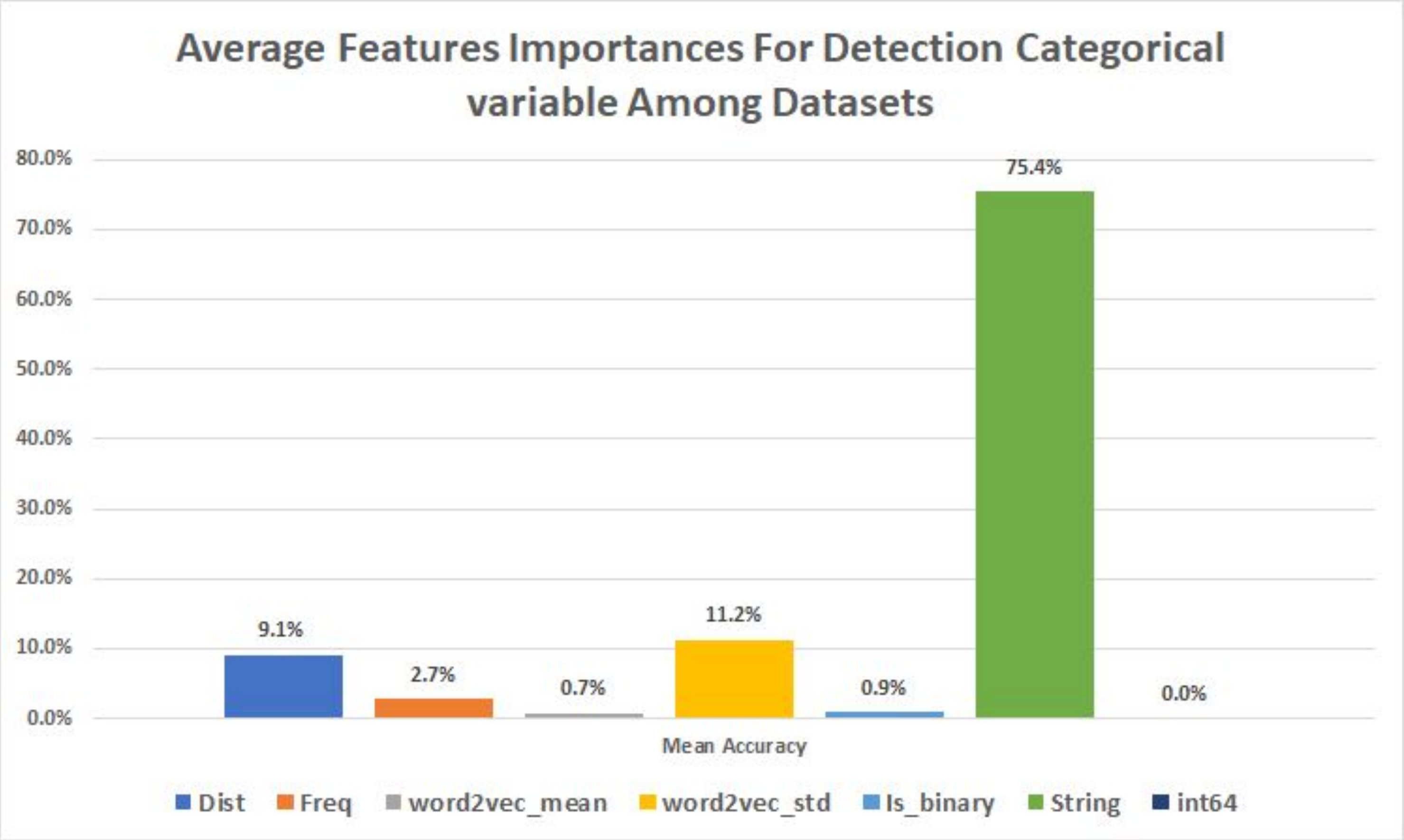
FIRST PHASE (Backup) EVALUATION

List of False Negative columns				
True Labels	**Data Set Name**	**Column Name**	**Prediceted Labels**	**Unique Values**
Nominal	Titanic	PassengerId	Numrical	1,2,3...
	Adult	race	Ordinal	Black, White, Other, Amer-Indian-Eskimo, Asian-Pac-Islander
	Bridges	RIVER	Ordinal	M, A, O, Y
	Nursery	form	Ordinal	complete, incomplete,completed, foster
	New_dataset	color	Ordinal	
	Heart	sex	Numrical	0,1
	Heart	target	Numrical	0,1
	Books	author_id	Numrical	1,2,3,4,5
	Books	score	Numrical	0,1,2,3
Ordinal	Titanic	Pclass	Nominal	1,2,3
	Car	price	Nominal	low,medium,larg
	Adult	education	Nominal	11th,HS-grad, Some-college, 10th, Prof-school, ...etc
	Adult	education-num	Numrical	1,2,3,4,5,6,7
	Audiology	air	Nominal	moderate, severe, normal
	Audiology	speech	Nominal	normal, good, perfect, bad, poor, unmeasured
	Random	Size	Nominal	Small,Large
	Nursery	housing	Nominal	convenient,less_conv, critical
	Nursery	social	Nominal	nonprob,slightly_prob, problematic

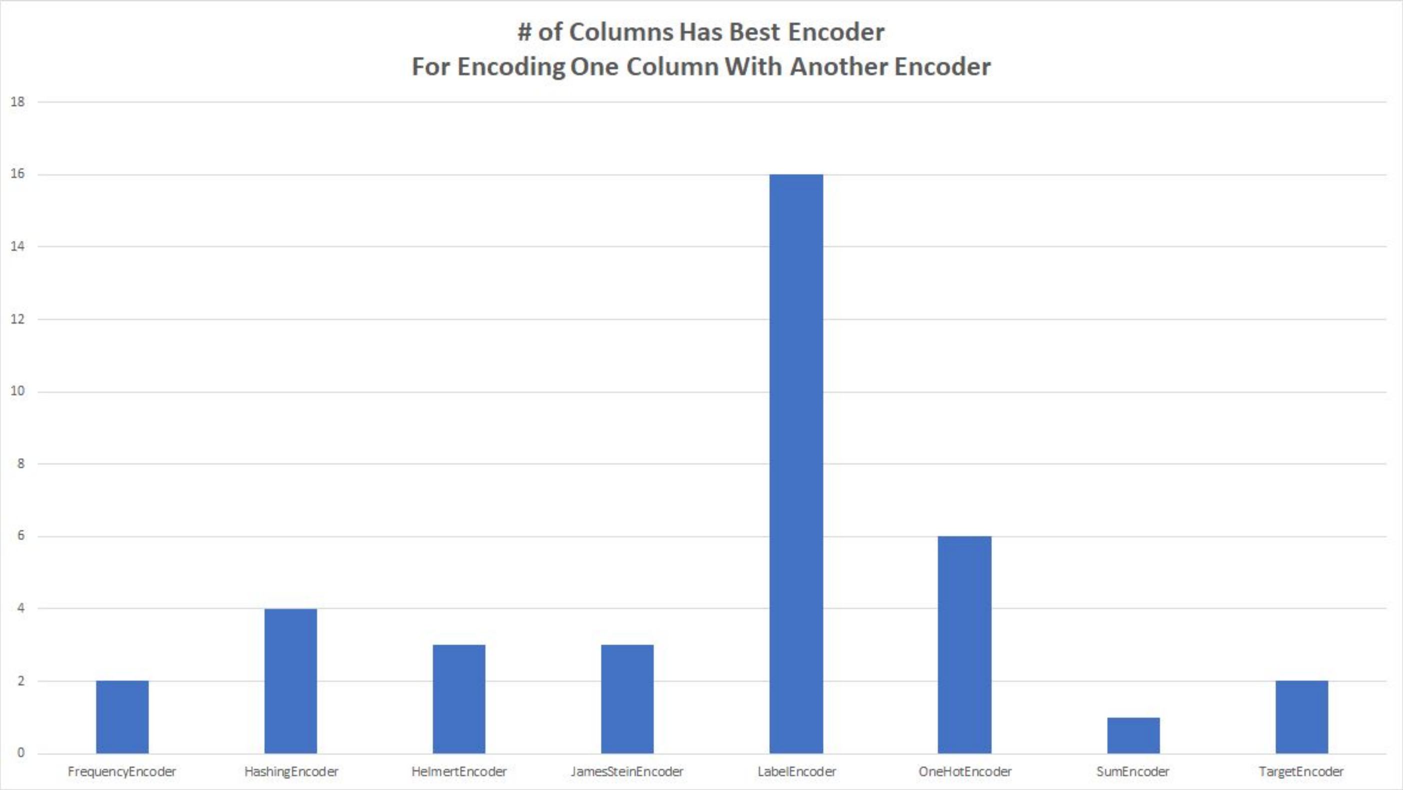
SECOND PHASE (Backup) ENCODERS

Encoder	Description
Helmert Encoder	The mean of the dependent variable for a level is compared to the mean of the dependent variable over all previous levels.
Sum Encoder	The mean of the dependent variable for a given level to the overall mean of the dependent variable over all the levels.
JamesStein Encoder	Is a biased estimator of the mean of Gaussian random vectors. It can be shown that the James–Stein estimator dominates the "ordinary" least square approach.
Hashing Encoder	Generating a hash value for each data-value. Some info loss due to collisions.
Frequency Encoder	Replacing each data-value by its frequency.
Binary Encoder	Converting each data-value to binary digits. Each binary digit gets one column. Some info loss but fewer dimensions.
Target Encoder	Is the process of replacing a data-value by the mean of the target variable.
BaseN Encoder	Base-N encoder encodes the categories into arrays of their base-N representation.

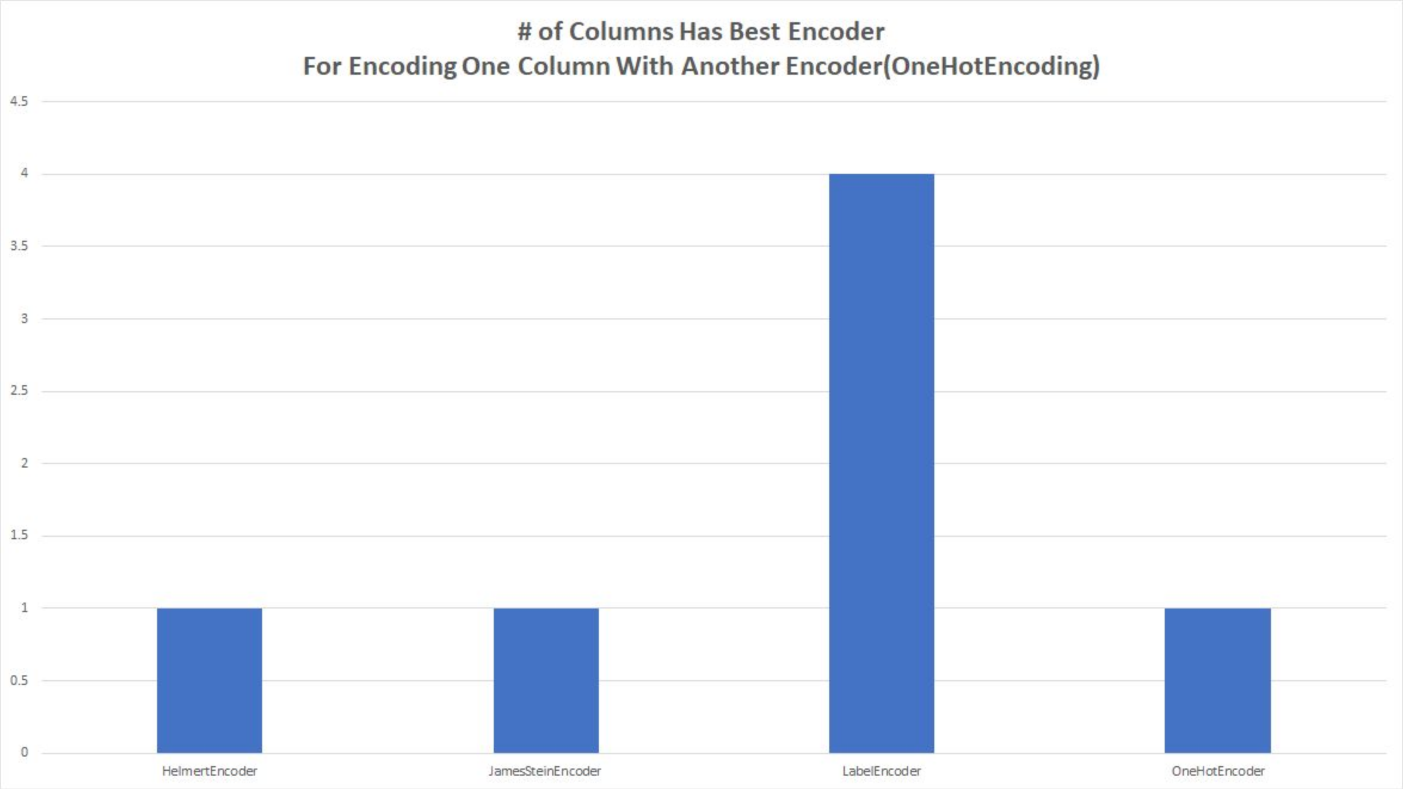
FIRST PHASE(BACKUP) RESULTS



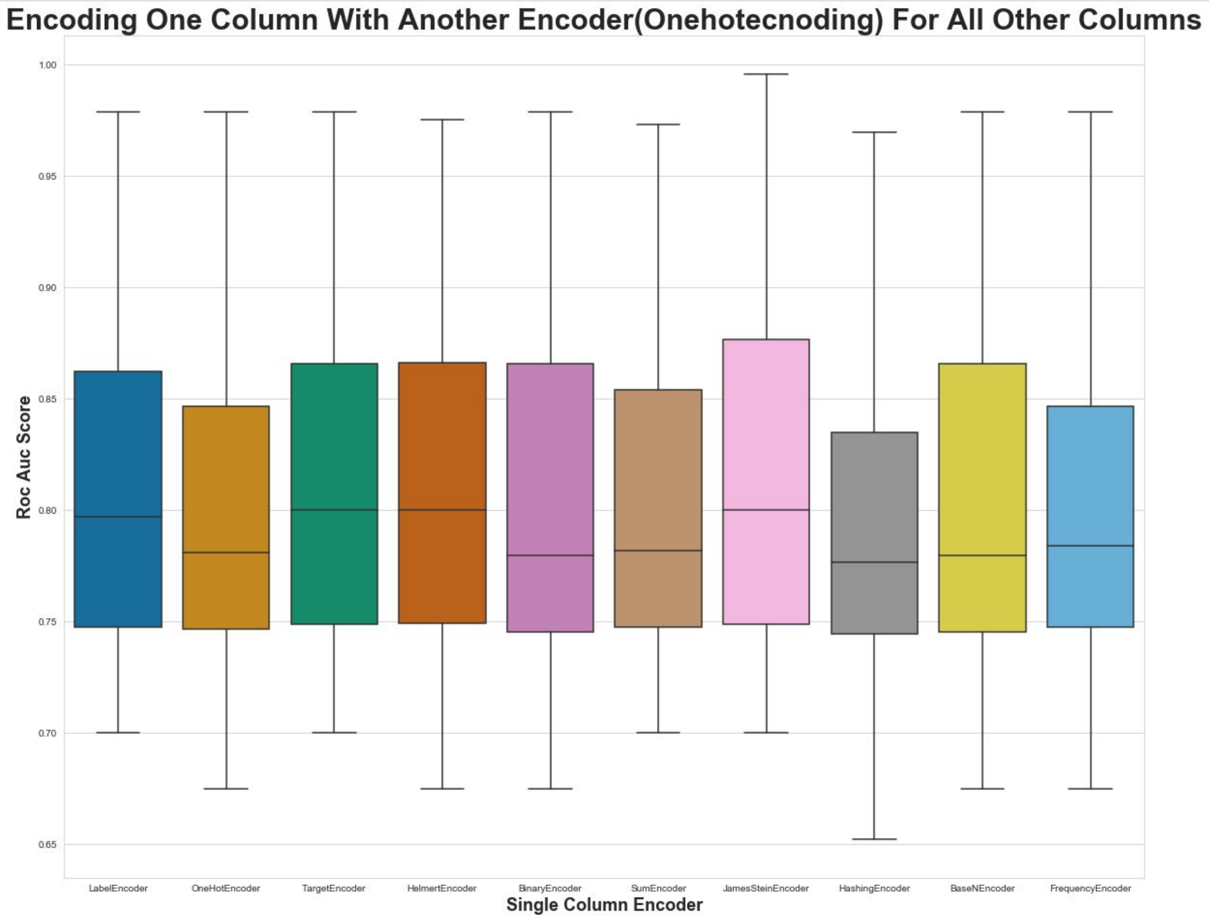
SECOND PHASE (Backup) RESULTS



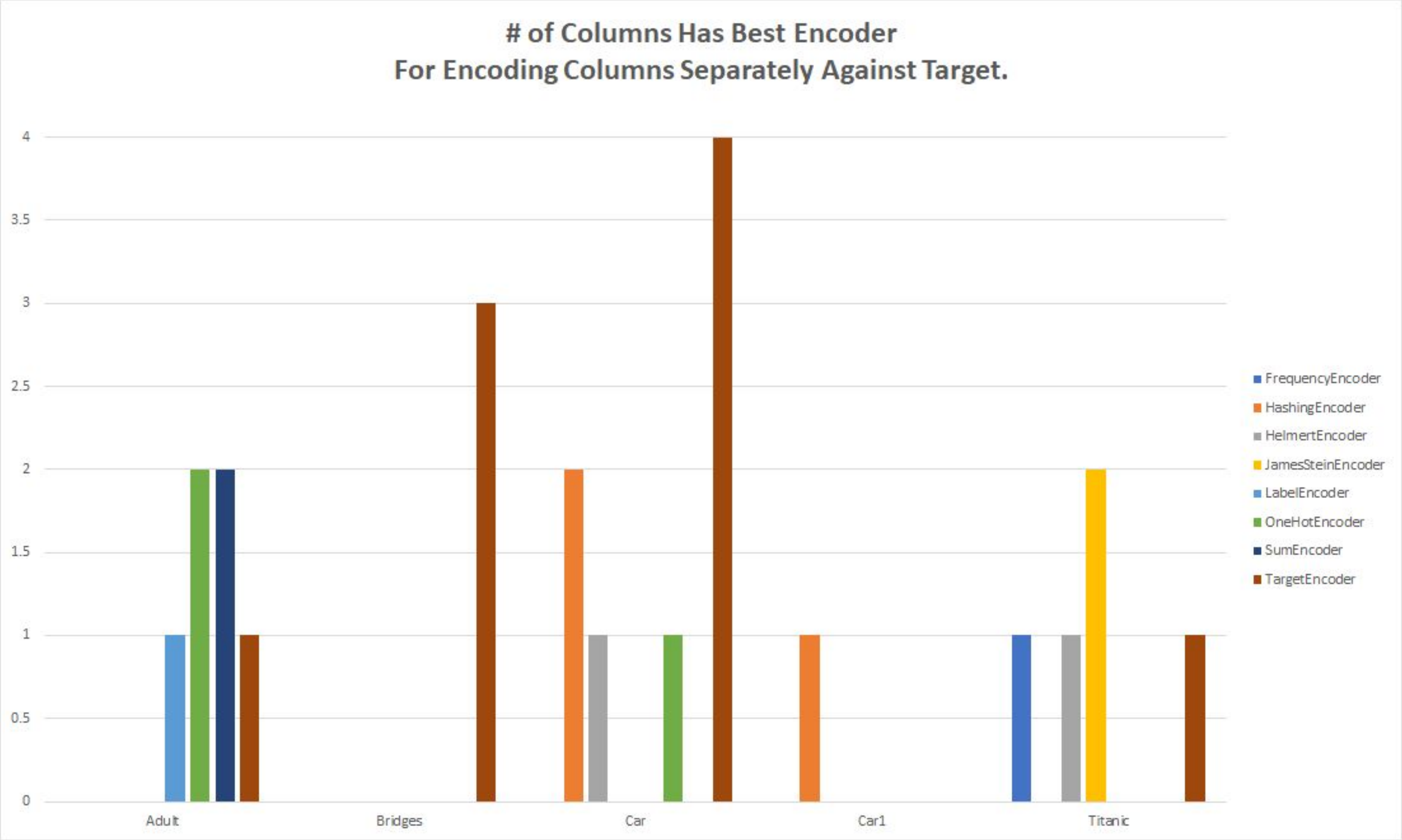
SECOND PHASE (Backup) RESULTS



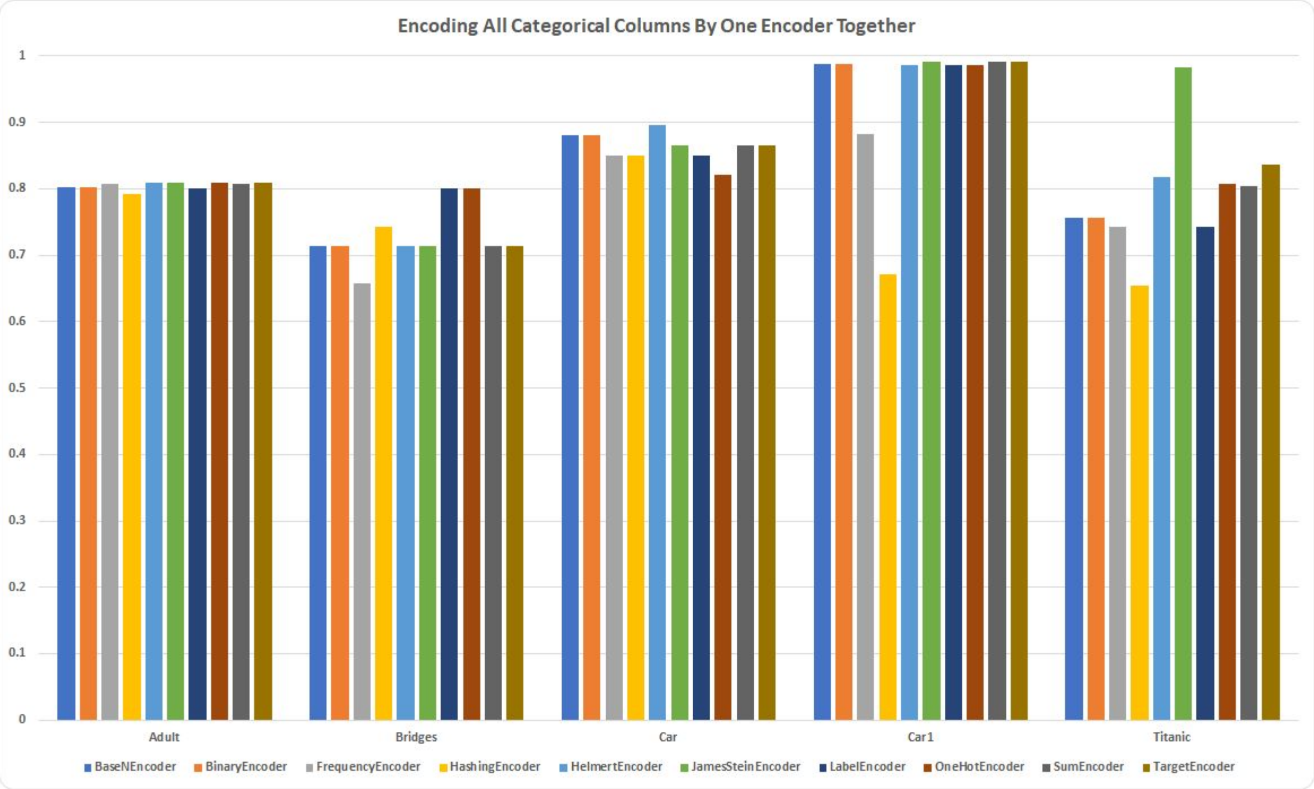
SECOND PHASE (Backup) RESULTS



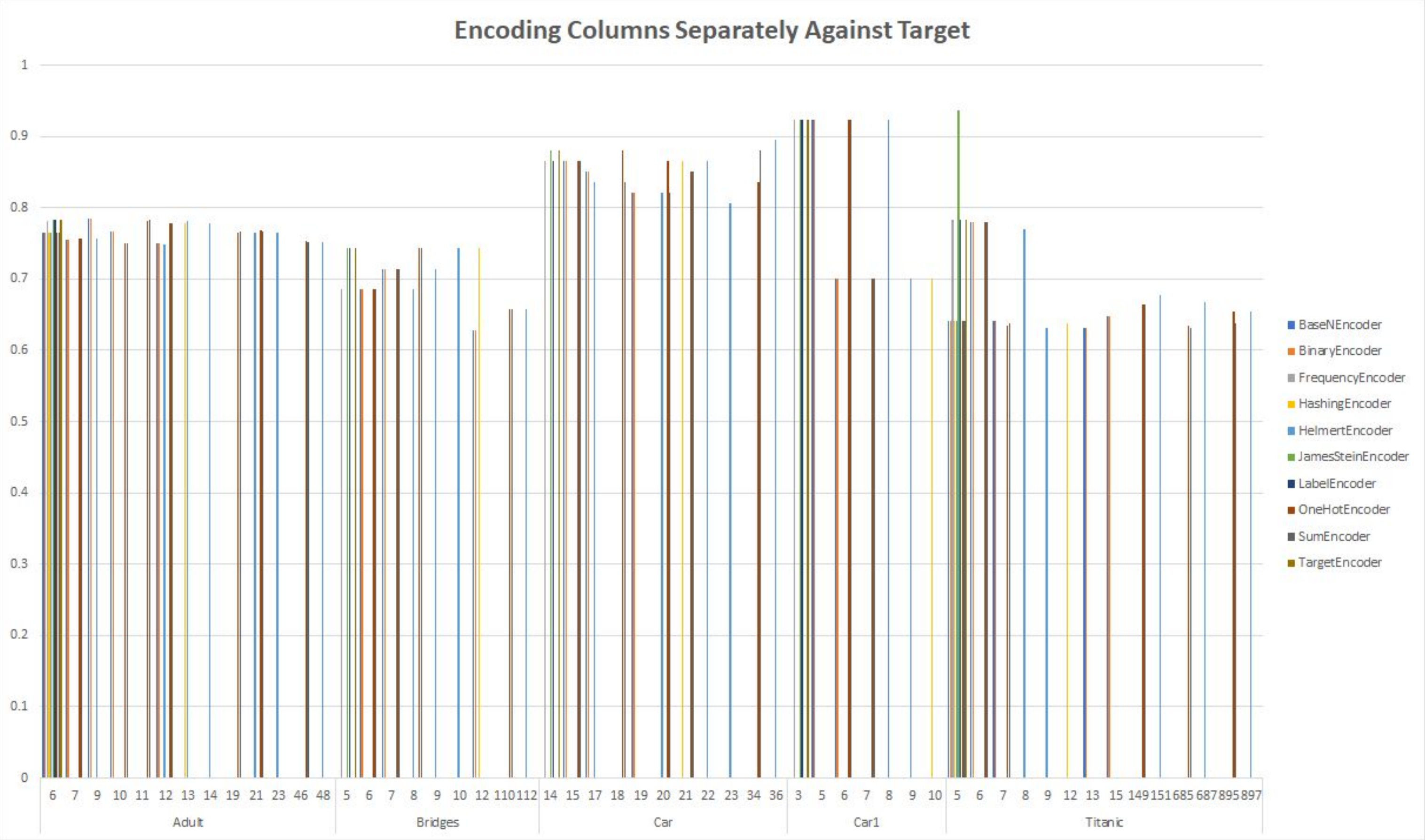
SECOND PHASE (Backup) RESULTS



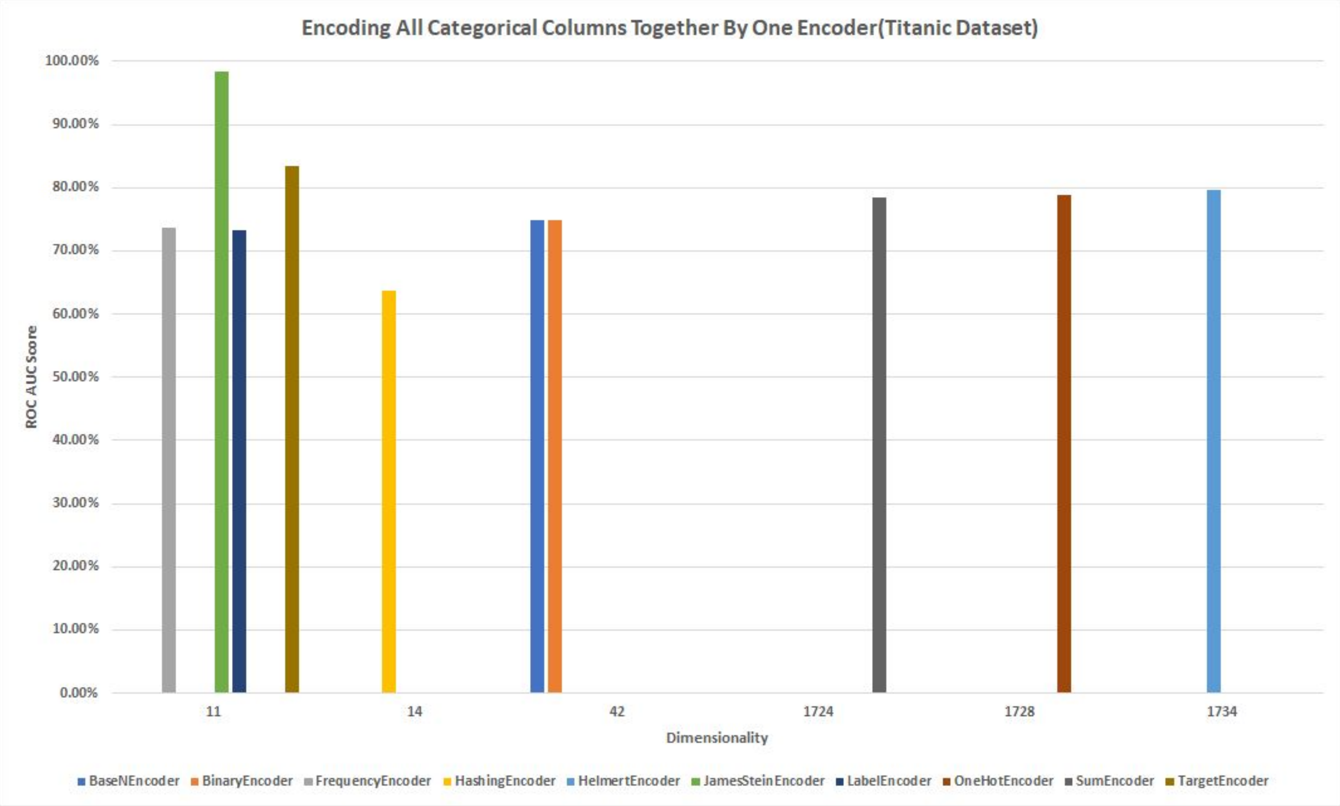
SECOND PHASE (Backup) RESULTS



SECOND PHASE (Backup) RESULTS



What will happen if we encode all categorical features together by the exact same encoder? (Backup)



- ★ JamesStein encoder achieved high score & low dimensionality.
- ★ Sum encoder, One hot encoder and Helmert encoder produced high dimensionality.
- ★ The rest encoders produced low dimensionality.

SECOND PHASE (Backup) RESULTS