Using Target Encoding to Improve Model Predictions

Jo-fai (Joe) Chow
Data Science Evangelist, H2O.ai
joe@h2o.ai / @matlabulous

Download: bit.ly/h2o_meetups

ABOUT ME

Jo-fai Chow

Data Science Evangelist & Community Manager at H2O.ai

United Kingdom

Add profile section | More...

H2O.

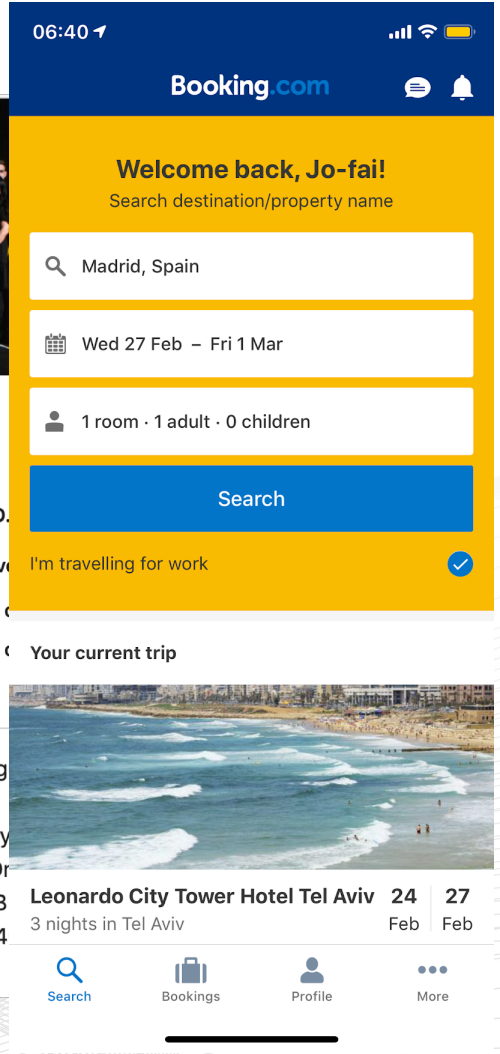
Univ

See

See

Jo-fai (or Joe) has multiple roles at H2O.ai. He is best known as the #360Selfie g

On LinkedIn, he is the data science evangelist and community manager but every
photography skills totally overshadow his data science knowledge these days. Or
like a die-hard MATLAB fanboy with the handle @matlabulous (because MATLAB
at Uni). Since joining H2O.ai in 2016, Joe has delivered H2O talks/workshops in 4
Europe and US.



About H2O.ai

H2O.ai is the open source leader in AI

H2O.ai is a visionary Silicon Valley open source software company that created and reimagined what is possible. We are a company of makers that brought to market new platforms and technologies to drive the AI movement. We are the makers of, H2O, the leading open source data science and machine learning platform used by nearly half of the Fortune 500 and trusted by over 14,000 organizations and hundreds of thousands of data scientists around the world.

Our approach is to be open, transparent and push the bleeding edge. Our philosophy is to create a culture of makers: community, customers, partners, entrepreneurs and our own "makers gonna make". Our vision is to democratize AI for everyone. Not just a select few. We enable this with our award winning, H2O Driverless AI, the platform that uses AI to do AI to make it easier, faster and cheaper to deliver expert data science as a force multiplier for every enterprise. We want everyone to explore, learn, dream and imagine a new future.

222 OF FORTUNE 500 ♥ H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES

H_2O.ai

14,000 Companies using H2O

H2O World in San Francisco

4th & 5th Feb, 2019

Our most-attended event to date

Attendance: 900

H2O.ai Product Suite

H2O.ai

H2O	Spark + H2O SPARKLING WATER	H2O4GPU
In-memory, distributed machine learning algorithms with H2O Flow GUI	H2O AI open source engine integration with Spark	Lightning fast machine learning on GPUs

Open Source

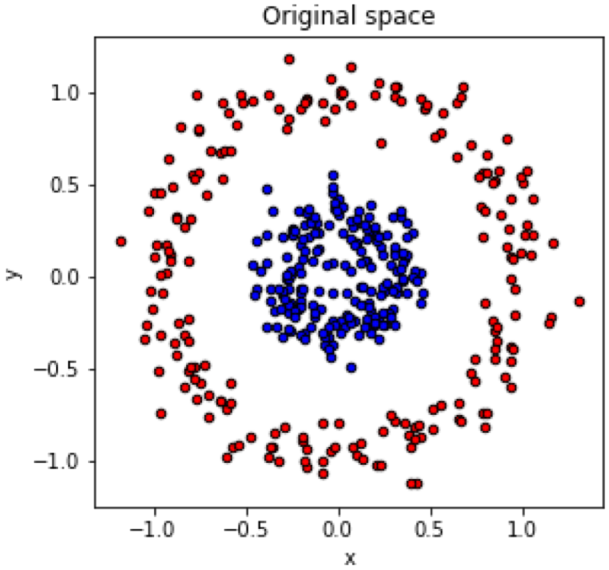
- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise support subscriptions

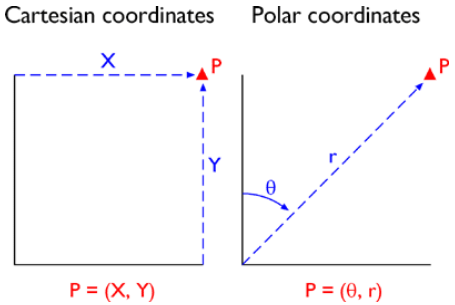
DRIVERLESSAI

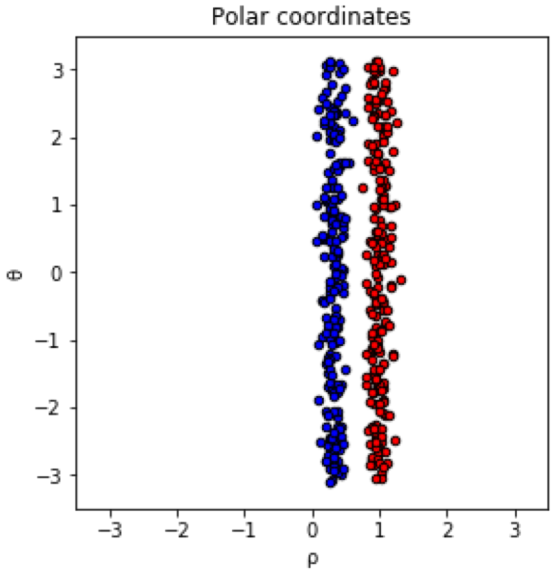
Automatic feature engineering, machine learning and interpretability

- Enterprise software
- Built for domain users, analysts and data scientists – GUI-based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

Why Feature Engineering?







$$\begin{bmatrix} \mathbf{X} \end{bmatrix} \dashrightarrow \begin{bmatrix} \mathbf{U} \end{bmatrix}$$

Feature Engineering: Target Mean Encoding

H2O.ai

What?

- Replace categorical variables with the mean of the response

Why?

- Categorical variables increase the number of features (dummy encoding) and can cause us to overfit

Caution!

- If it is applied without care, it may lead to overfitting

Target Mean Encoding

H2O.ai

Categorical Feature	Target	Target Mean Encoding
A	0	
A	0	
A	0	
B	1	
B	0	
B	0	
C	1	

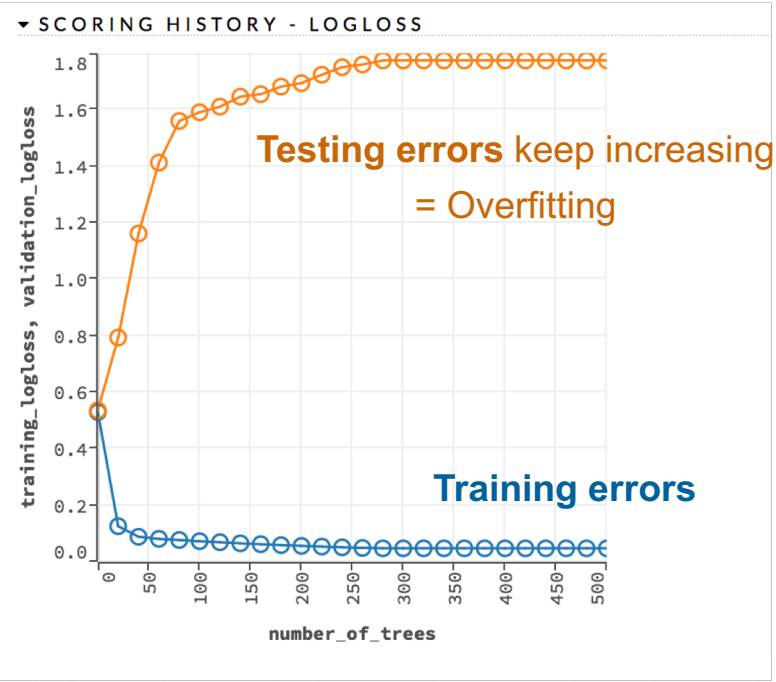
Target Mean Encoding Done Wrong

H2O.ai

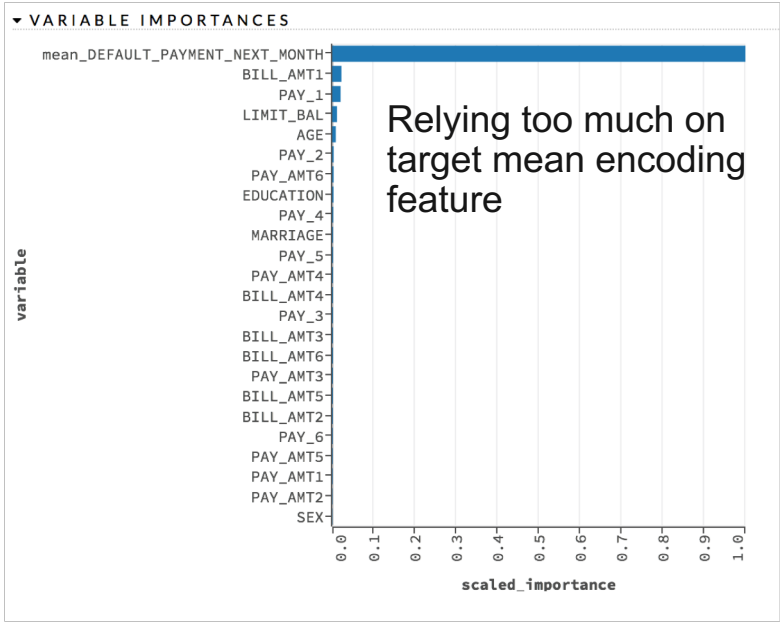
Categorical Feature	Target		Target Mean Encoding
A	0		0
A	0	mean	0
A	0		0
B	1		0.33
B	0	mean	0.33
B	0		0.33
C	1	same value →	1

***Worst Case Scenario**: Response Column = Mean Target Encoding*

Target Mean Encoding Done Wrong



Scoring History: Training vs Testing



Data Leakage Feature is the only important feature

Solution: Cross-Validation Target Encoding

H2O.ai

Fold	Categorical Feature	Target
1	A	0
2	A	0
3	A	0
2	B	1
1	B	0
3	B	0
1	C	1

Solution: Cross-Validation Target Encoding

H2O.ai

Fold	Categorical Feature	Target
2	A	0
3	A	0
2	B	1
3	B	0

Fold	Categorical Feature	Target
1	A	0
1	B	0
1	C	1

Solution: Cross-Validation Target Encoding

H2O.ai

Fold	Categorical Feature	Target	Target Mean Encoding
2	A	0	0
3	A	0	0
2	B	1	0.5
3	B	0	0.5

Fold	Categorical Feature	Target
1	A	0
1	B	0
1	C	1

Solution: Cross-Validation Target Encoding

H2O.ai

Fold	Categorical Feature	Target	Target Mean Encoding
2	A	0	0
3	A	0	0
2	B	1	0.5
3	B	0	0.5

Fold	Categorical Feature	Target	CV Target Encoding
1	A	0	**0**
1	B	0	
1	C	1	

Solution: Cross-Validation Target Encoding

H2O.ai

Fold	Categorical Feature	Target	Target Mean Encoding
2	A	0	0
3	A	0	0
2	B	1	0.5
3	B	0	0.5

Fold	Categorical Feature	Target	CV Target Encoding
1	A	0	0
1	B	0	**0.5**
1	C	1	

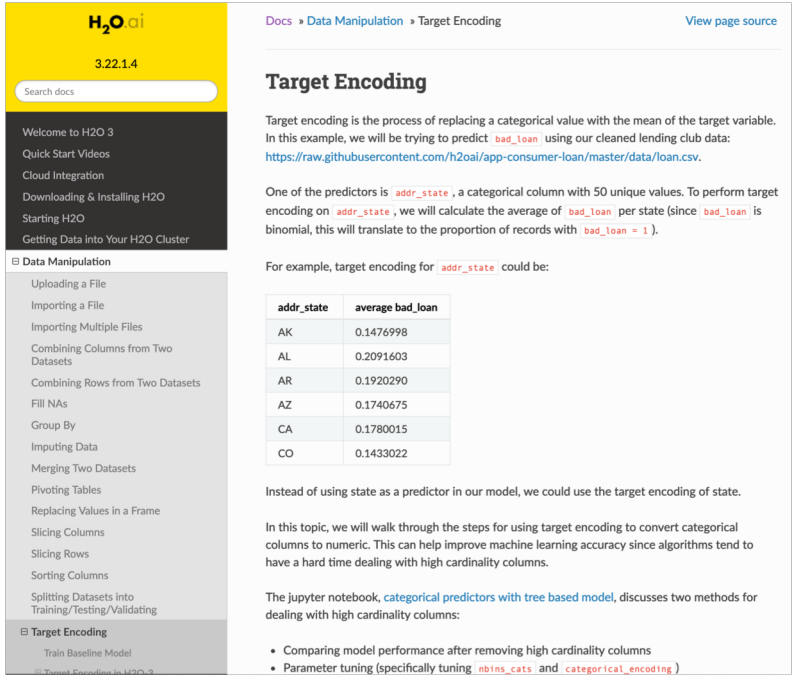
Solution: Cross-Validation Target Encoding

Fold	Categorical Feature	Target	Target Mean Encoding
2	A	0	0
3	A	0	0
2	B	1	0.5
3	B	0	0.5

No info.

Fold	Categorical Feature	Target	CV Target Encoding
1	A	0	0
1	B	0	0.5
1	C	1	NA

Using CV Target Encoding with H2O



H2O.ai

3.22.1.4

Search docs

Welcome to H2O 3
Quick Start Videos
Cloud Integration
Downloading & Installing H2O
Starting H2O
Getting Data into Your H2O Cluster
Data Manipulation
Uploading a File
Importing a File
Importing Multiple Files
Combining Columns from Two Datasets
Combining Rows from Two Datasets
Fill NAs
Group By
Imputing Data
Merging Two Datasets
Pivoting Tables
Replacing Values in a Frame
Slicing Columns
Slicing Rows
Sorting Columns
Splitting Datasets into Training/Testing/Validating
Target Encoding
Train Baseline Model

Docs » Data Manipulation » Target Encoding View page source

Target Encoding

Target encoding is the process of replacing a categorical value with the mean of the target variable. In this example, we will be trying to predict `bad_loan` using our cleaned lending club data: https://raw.githubusercontent.com/h2oai/app-consumer-loan/master/data/loan.csv.

One of the predictors is `addr_state`, a categorical column with 50 unique values. To perform target encoding on `addr_state`, we will calculate the average of `bad_loan` per state (since `bad_loan` is binomial, this will translate to the proportion of records with `bad_loan = 1`).

For example, target encoding for `addr_state` could be:

addr_state	average bad_loan
AK	0.1476998
AL	0.2091603
AR	0.1920290
AZ	0.1740675
CA	0.1780015
CO	0.1433022

Instead of using state as a predictor in our model, we could use the target encoding of state.

In this topic, we will walk through the steps for using target encoding to convert categorical columns to numeric. This can help improve machine learning accuracy since algorithms tend to have a hard time dealing with high cardinality columns.

The jupyter notebook, categorical predictors with tree based model, discusses two methods for dealing with high cardinality columns:

- Comparing model performance after removing high cardinality columns
- Parameter tuning (specifically tuning `nbins_cats` and `categorical_encoding`)

Comparison

Let's compare all three models:

Evaluation (AUC, Higher = Better):

5-Fold CV (with addr_state): `0.7045098` vs. (without addr_state): `0.7061583` vs. (with TE): `0.7072099`

Test (with addr_state): `0.7069701` vs. (without addr_state): `0.7076197` vs. (with TE) `0.708911`

Higher AUC with TE

My code example:

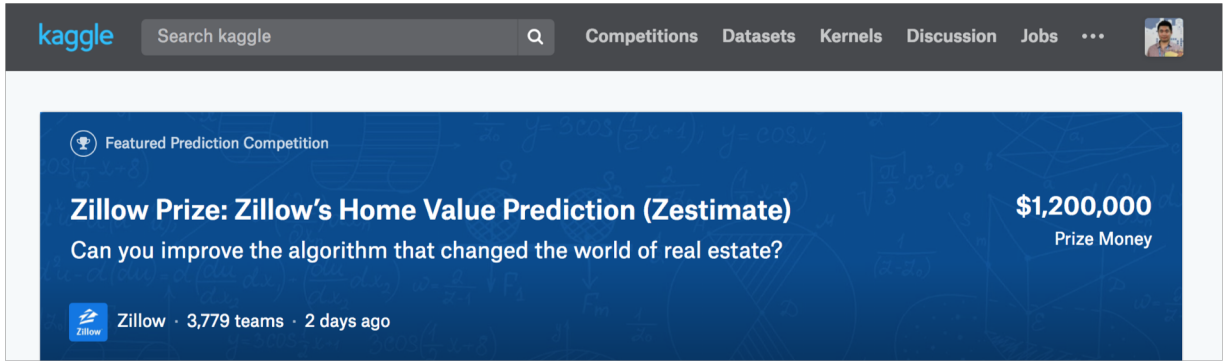
https://github.com/h2oai/h2o-meetups/tree/master/2019_02_26_Tel_Aviv

H2O-3 Documentation:

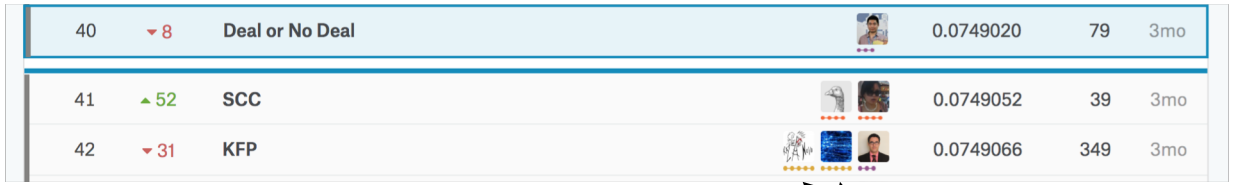
http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-munging/target-encoding.html

CV Target Encoding + Other Feature Engineering Tricks

Competition Round One (Top 100 to Next Round)



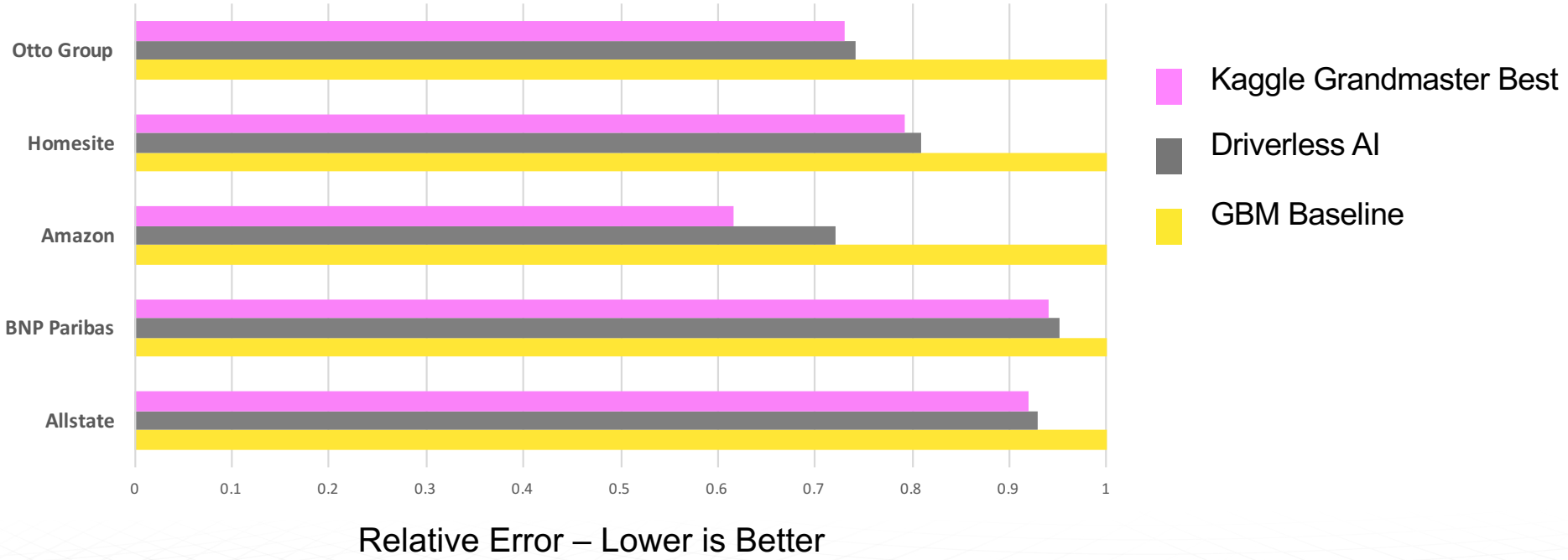
40 out of 3779 teams



Finished above my H2O Kaggle Grandmasters colleagues ☺

Appendix

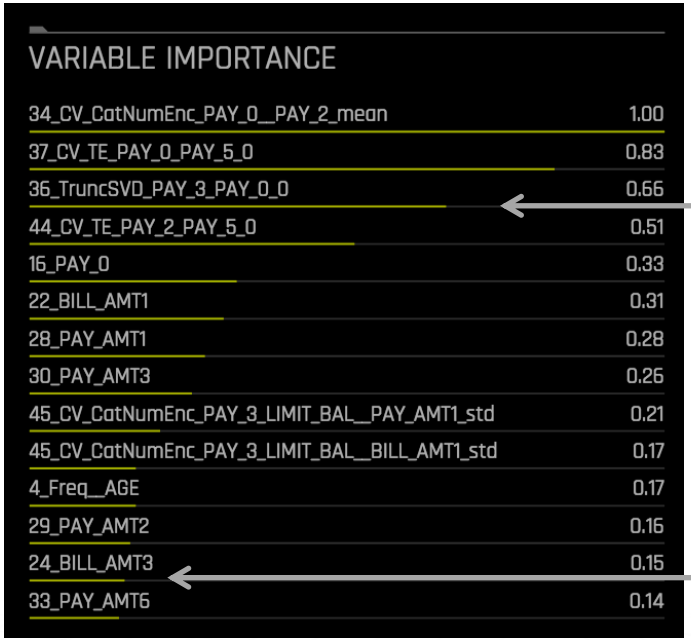
How Does Feature Engineering Effect Accuracy?



Kaggle Grandmaster Out-of-the-Box

Feature Transformations

- Automatic text handling
- Frequency encoding
- Cross validation target encoding
- Truncated singular value decompression
- Clustering and more



Accuracy

- Automatic feature engineering to increase accuracy - AlphaGo for AI
- Automatic Kaggle Grandmaster recipes in a box for solving wide variety of use-cases
- Automatic machine learning to find and tune the right ensemble of models