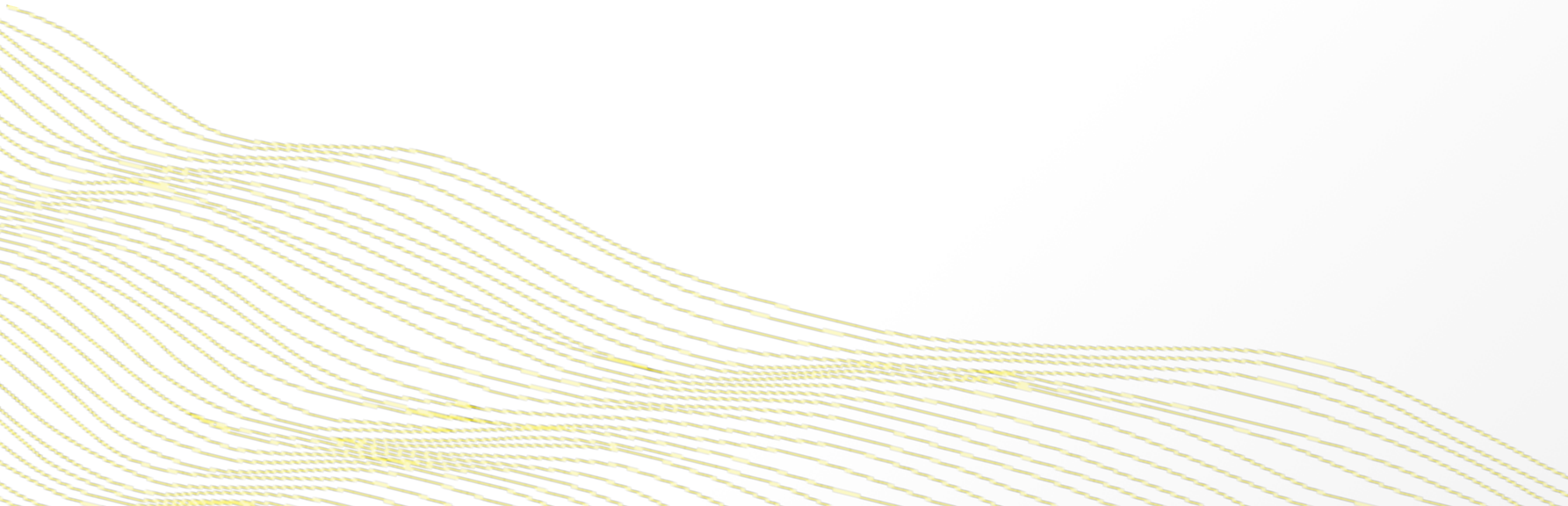
Driverless AI

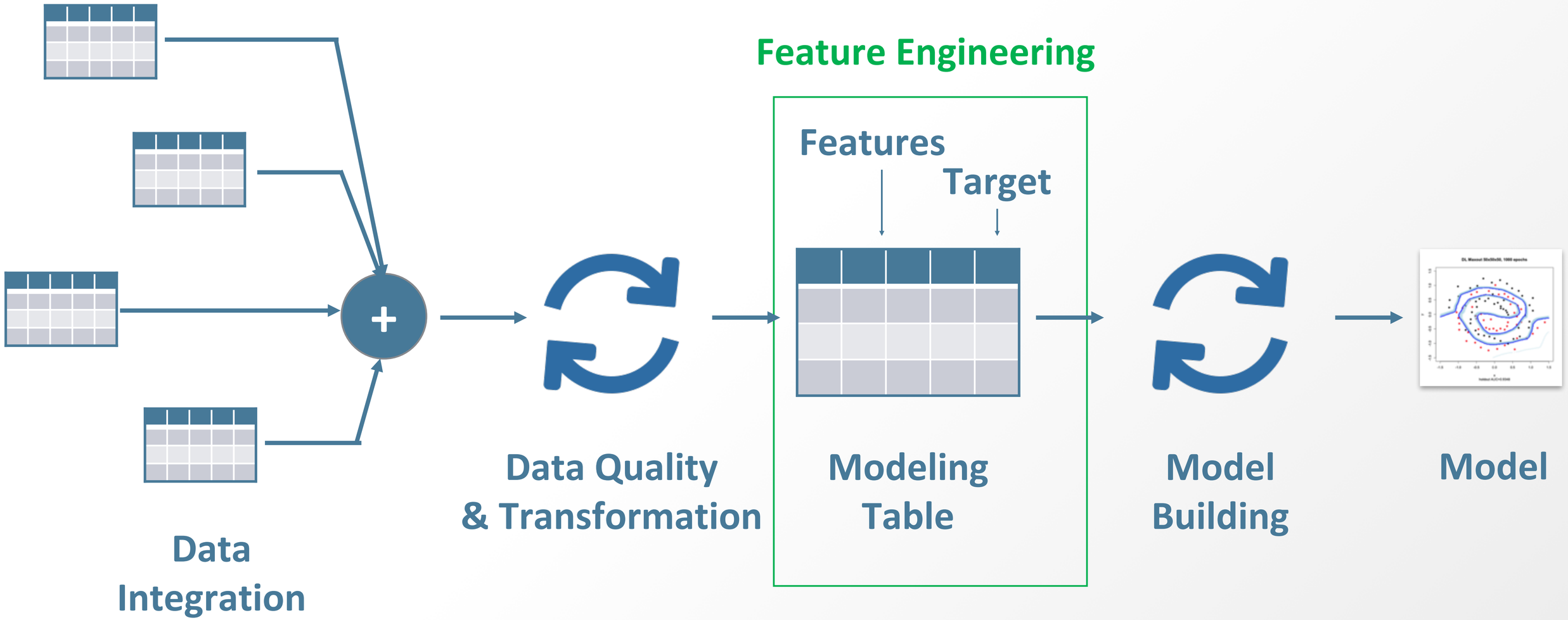
Feature Engineering

Megan Kurka

Feature Engineering



Typical Enterprise Machine Learning Workflow



What is Feature Engineering?

Sales Forecasting

Store	Date	Sales
1	2015-07-31	$5,263
1	2015-08-01	$6,064
2	2015-07-29	$15,344

Mean Absolute Percent Error = 37%

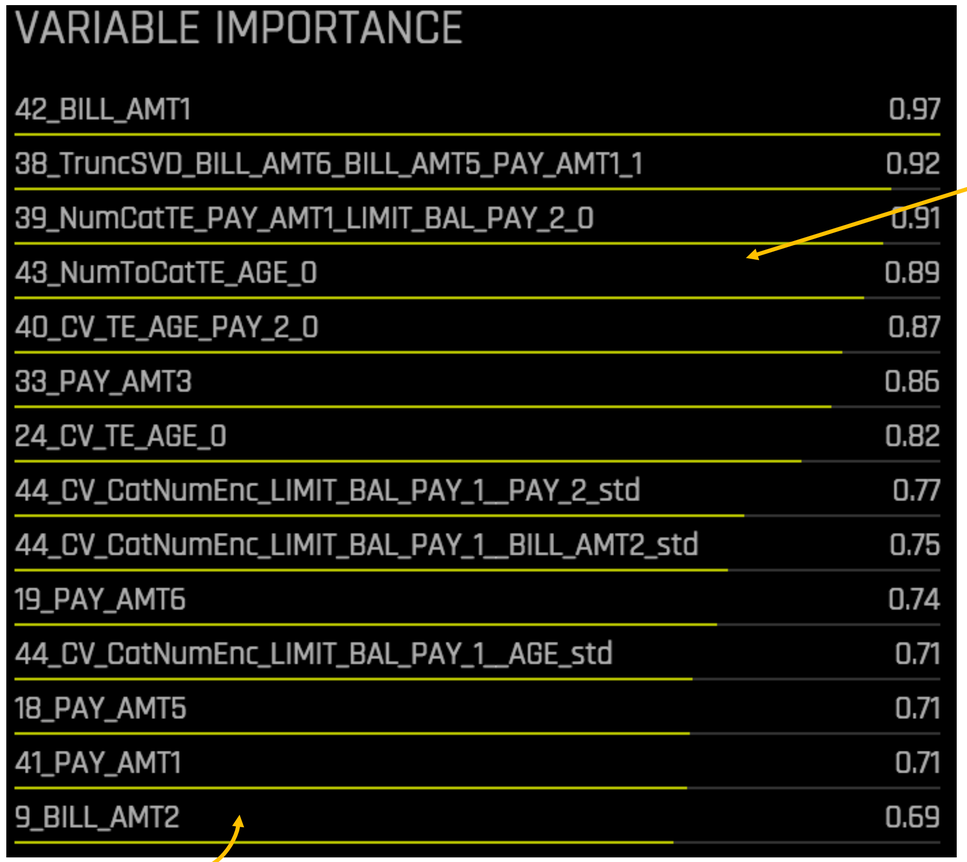
What is Feature Engineering?

Sales Forecasting

Store	Month	Day Of Week	Sales Previous Day	Sales
1	July	Friday	$5,431	$5,263
1	August	Saturday	$5,102	$6,064
2	July	Wednesday	$14,958	$15,344

Mean Absolute Percent Error = 16%

Auto Feature Generation
Kaggle Grand Master Out of the Box



Generated Features

Original Features

Feature Transformations

- Cross Validation Categorical Encoding
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD
- Clustering and more

How does Feature Engineering Effect Accuracy?

- Results for Driverless AI are shown using a single, un-tuned model
- Similar error to a Kaggle Grandmaster from solely feature engineering

Target Mean Encoding

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

Target Mean Encoding

PAY_1	DEFAULT PAYMENT
Up To Date	0
Up To Date	0
Up To Date	0
Missed 1 Mo	1
Missed 1 Mo	0
Missed 1 Mo	0
Missed 5 Mo	1

- Customers who are up to date on their payments
- No occurrence of Default

- Customers who are one month behind on their payments
- Some occurrence of Default

- Customers who are five months behind on their payments
- Always occurrence of Default

Target Mean Encoding

PAY_1	DEFAULT PAYMENT	Mean Target Encoding
Up To Date	0	0
Up To Date	0	0
Up To Date	0	0
Missed 1 Mo	1	0.33
Missed 1 Mo	0	0.33
Missed 1 Mo	0	0.33
Missed 5 Mo	1	1

Data Leakage

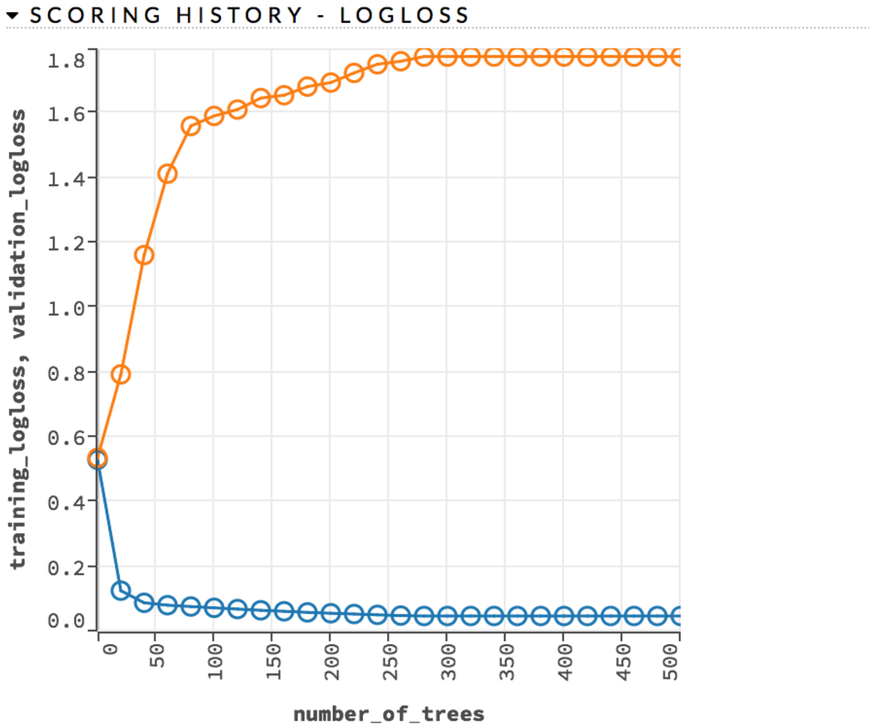
- Mean Target Encoding is based on the response column of the rows
- The lower the number of rows in the group, the more it reveals the response column value

PAY_1	DEFAULT PAYMENT	Mean Target Encoding
Missed 5 Mo	1	1

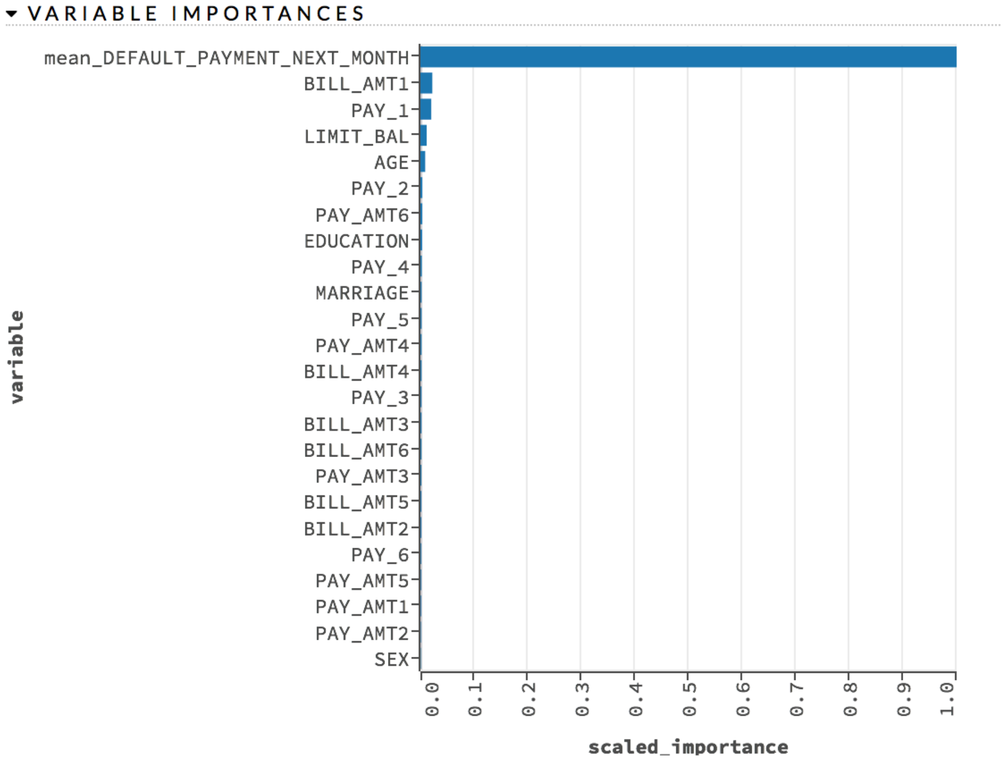
Worst Case Scenario: Response Column = Mean Target Encoding

Overfitting

- Data Leakage causes overfitting
- Great results on the training, bad results on hold out data



Scoring History: Training vs Testing



Data Leakage Feature is the only important feature

H2O.ai

Known Issues

1. Anything we don't know when scoring, shouldn't be used as training features
   - Mean Target Encoding incorporates the response value
   - We would not know the response value when scoring

2. The less frequent the categorical value is, the more the mean target encoding is overfitting

PAY_1	DEFAULT PAYMENT	Mean Target Encoding
Missed 5 Mo	1	1

Worst Case Scenario: Response Column = Mean Target Encoding

Cross Validation Target Encoding

- Use Cross Validation folds to calculate average Target
- The Mean Target Encoding for each row never uses that row's response

Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
1	Up to Date	0	0
2	Up to Date	0	0
3	Up to Date	0	0
2	Missed 1 Mo	1	0.33
1	Missed 1 Mo	0	0.33
3	Missed 1 Mo	0	0.33
1	Missed 5 Mo	1	1

Cross Validation Target Encoding

For Fold 1, we will calculate Mean Target Encoding using all rows not in Fold 1

Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
2	Up to Date	0	0
3	Up to Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

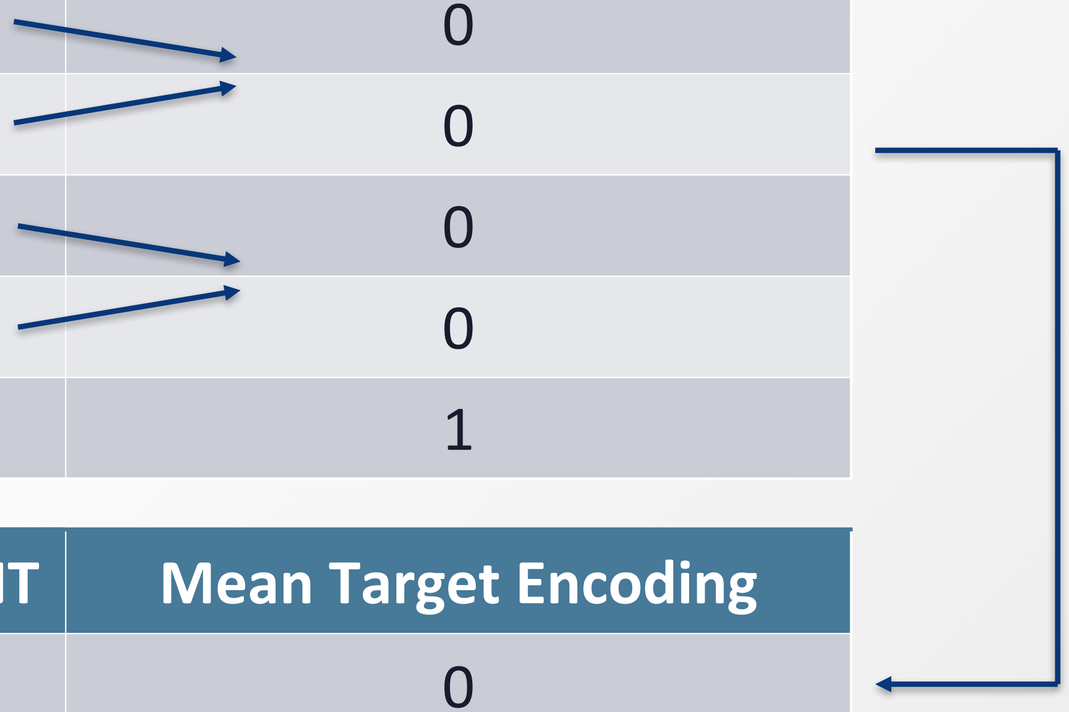
Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
1	Up to Date	0	0
1	Missed 1 Mo	1	0.5
1	Missed 5 Mo	1	NA

Cross Validation Target Encoding

For Fold 2, we will calculate Mean Target Encoding using all rows not in Fold 1

Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
1	Up to Date	0	0
3	Up to Date	0	0
1	Missed 1 Mo	0	0
3	Missed 1 Mo	0	0
1	Missed 5 Mo	1	1

Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
2	Up to Date	0	0
2	Missed 1 Mo	1	0



Cross Validation Target Encoding

- Mean Target Encoding was originally unique to each level
- Now Mean Target Encoding unique to each level and fold

Fold	PAY_1	DEFAULT PAYMENT	Mean Target Encoding
1	Up to Date	0	0
2	Up to Date	0	0
3	Up to Date	0	0
2	Missed 1 Mo	1	0
1	Missed 1 Mo	0	0.5
3	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	NA

Weighted Target Encoding

- Calculate Weighted Mean instead of Mean:

$$weight_1 * mean(level) + weight_2 * mean(dataset)$$

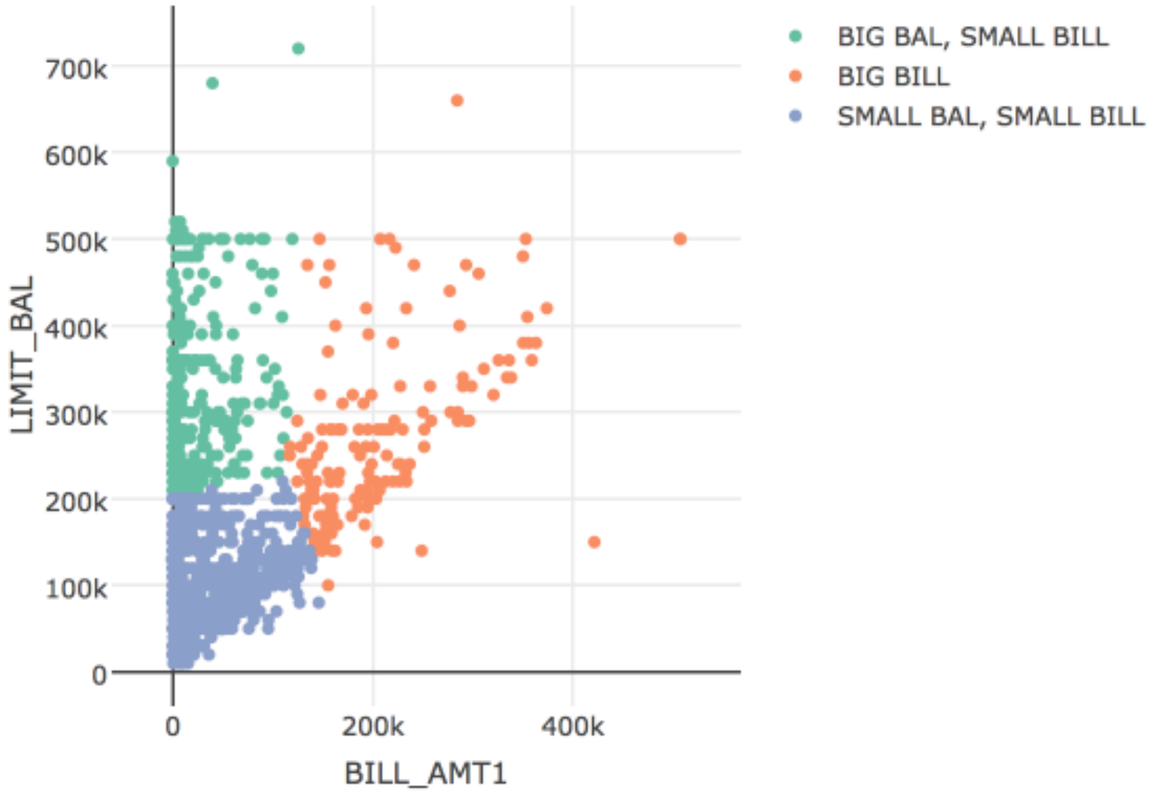
- The less frequent the level is, the greater the weight of the overall mean is

PAY_1	Count	Weight of Overall Mean	Mean Target Encoding	Weighted Mean Target Encoding
-1	3	0.5	0	0.14
1	3	0.5	0.33	0.31
5	1	0.9	1	0.36

Feature Encoding – Numeric Features

- Binning categoricals
  - Using quantiles
  - Using histogram bins
- Clustering
  - Use Cluster ID’s as new feature
- Dimensionality Reduction Techniques
  - Truncated SVD

Clustering - KMeans



LIMIT_BAL	BILL_AMT1	CLUSTER
$340,000	$307,650	• Big Bill
$300,000	$12,752	• Big Balance • Small Bill
$140,000	$35,206	• Small Balance • Small Bill
$30,000	$17,341	• Small Balance • Small Bill
$50,000	$7,010	• Small Balance • Small Bill
$230,000	$2,4524	• Big Balance • Small Bill

H2O.ai

Dimensionality Reduction – Truncated SVD

What?

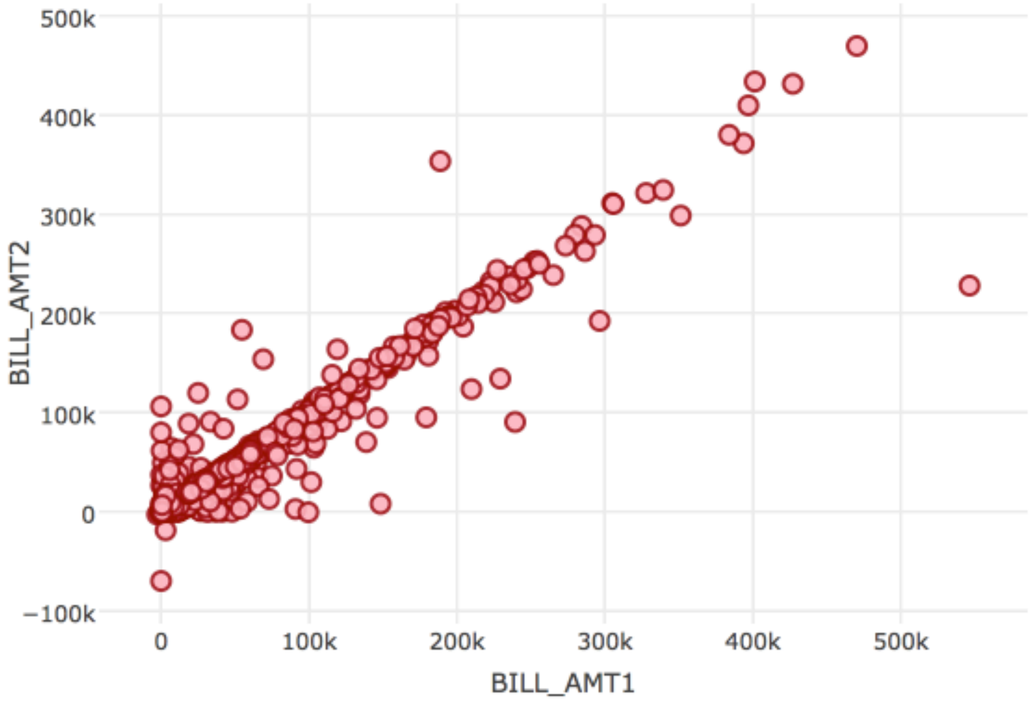
- Transform multiple features into a smaller number of dimensions

Why?

- Remove noise
- Remove multi-collinearity

Dimensionality Reduction – Truncated SVD

BILL_AMT1	BILL_AMT2
$948	$1,156
$51,377	$50,253
$351,232	$298,709
$3,706	$400,093
$0	$0
$166	$1,694



Dimensionality Reduction – Truncated SVD