Deep Feature Synthesis:
Towards Automating Data Science Endeavors

Xue Huanran 1701214297

2018.12.10

Introduction

Peking University

In this paper, the Data Science Machine is developed, which is able to derive predictive models from raw data automatically.

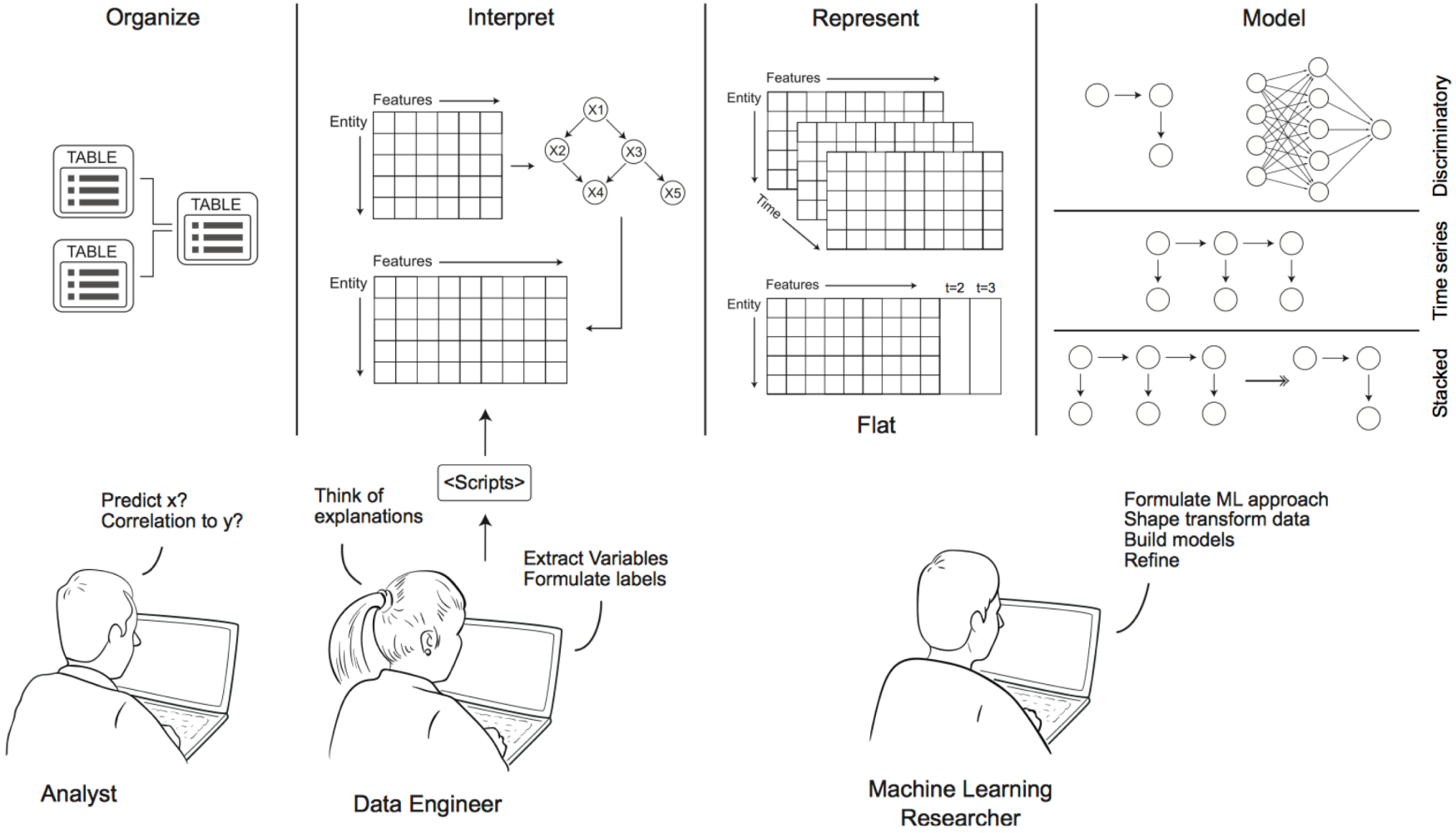
The pipeline of data science:

- Feature Engineering ---- Deep Feature Synthesis
- Select the predictive methodology
- Tuning the hyperparameters

Data Science Machine

A typical data science endeavor

Peking University



Contributions

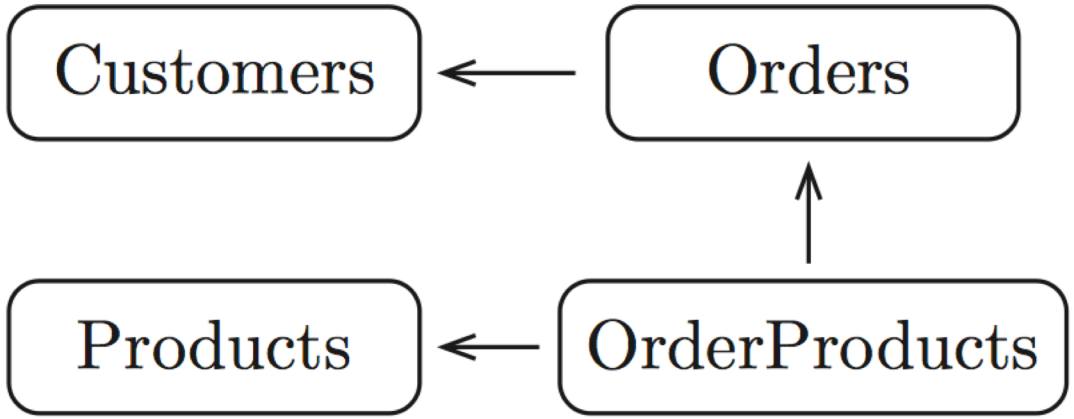
Peking University

- Automatically generate features via **Deep Feature Synthesis**.
- Autotune a machine learning pipeline to extract the most value out of the synthesized features.
- Produce submissions for online data science competitions.

Deep Feature Synthesis

Prototypical Problems and Motivation

Peking University



A simplified schema for an e-commerce website. There 4 entities. An arrow from one entity to another signifies that the first entity references the second in the databases.

Feature Synthesis Abstractions

Peking University

Given a dataset, we have entities given by $E^{1...K}$, where each entity table has $1 ... J$ features. Denote a specific entry as $x_{i,j}^{k}$, which is the value for feature j for the i^{th} instance of the k^{th} entity.

Next, given entities and their relationships, define a number of mathematical functions that are applied at two levels: at the **entity** level and at the **relational** level.

Two types of features

Peking University

Entity features (efeat): Entity features derive features by computing a value for each entry $x_{i,j}$.
LENGTH(), YEAR(), MONTH(), DAY(), WEEKDAY(),…

The second set of features is derived by jointly analyzing two related entities, E^l and E^k. These two entities relate to each other in one of two ways: **forward** and **backward**.

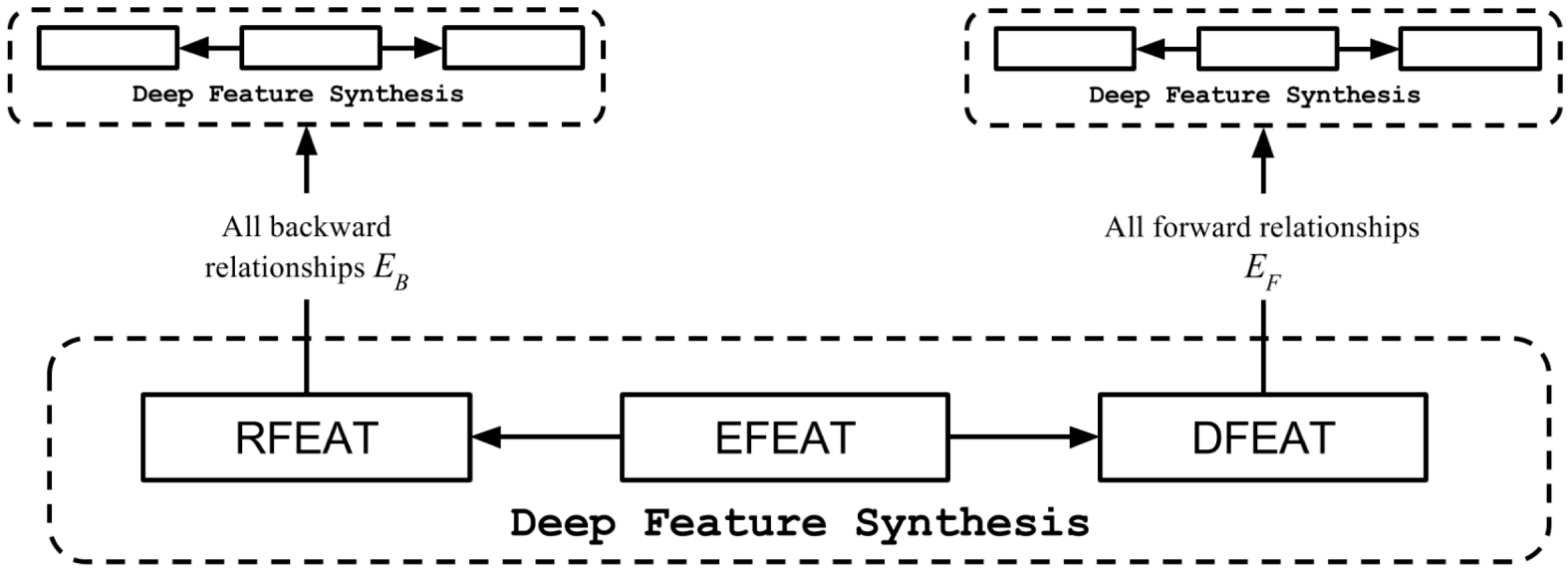
Forward (dfeat): Instance i has an explicit dependence on instance m. E.g. the **Orders** entity has a forward relationship with the **Customers**.

Backward (rfeat): The backward relation is the relationship from an instance i in E^k to all the instances $m = \{1 \dots M\}$ in E^l. E.g. the **Customers** entity has backward relationship with **Orders**.
AVG(), MAX(), MIN(), SUM(), STD(), COUNT()

How to generate features

Peking University

Consider a dataset of K entities, the goal is to extract **rfeat**, **dfeat** and **efeat** feaetures for a target E^k. Additionally, all the entities with which E^k has **forward** and **backward** relationships. These are denoted by sets E_F and E_B.



The recursive scenario of DFS

Peking University

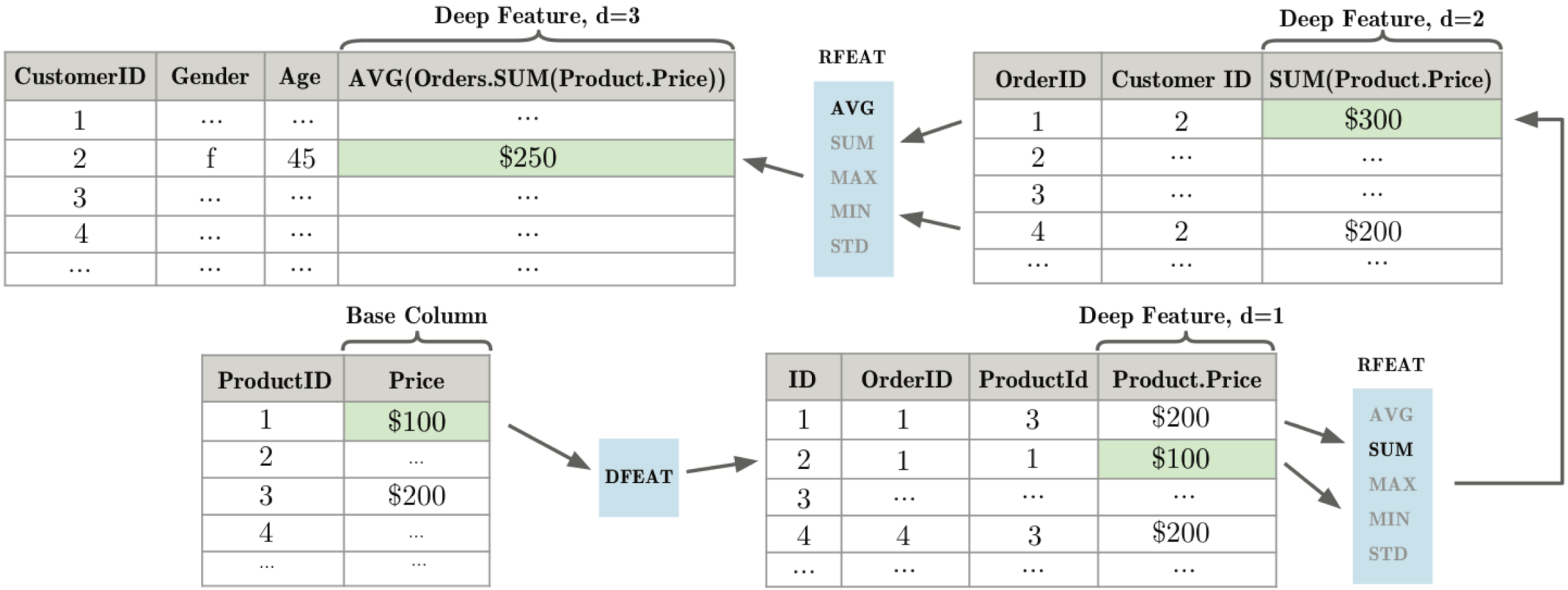
Algorithm 1 Generating features for target entity

function MAKE_FEATURES($E^i, E^{1:M}, E_V$)
    $E_V = E_V \cup E^i$
    $E_B =$ BACKWARD$(E^i, E^{1...M})$
    $E_F =$ FORWARD$(E^i, E^{1...M})$
    for $E^j \in E_B$ **do**
        MAKE_FEATURES$(E^j, E^{1...M}, E_V)$
        $F^j = F^j \cup$ RFEAT(E^i, E^j)
    for $E^j \in E_F$ **do**
        if $E^j \in E_V$ **then**
            CONTINUE
        MAKE_FEATURES$(E^j, E^{1...M}, E_V)$
        $F^i = F^i \cup$ DFEAT(E^i, E^j)
    $F^i = F^i \cup$ EFEAT(E^i)

The MAKE_FEATURES is presented to make features, F^i, for the i^{th} entity.
E_V keeps track of the entities that have been visited.

An Example of DFS

Peking University



Data Science Machine

Machine Learning Pipeline

Peking University

Data Science Machine implements a parametrized pipeline for data preprocessing, feature selection, dimensionality reduction, modeling and evaluation.

Data preprocessing: Remove the null variables, convert the categorical variables using one-hot encoding, normalizing the features.

Feature selection and dimensionality reduction: First, use truncated SVD transformation. Then, rank each feature by f-value with the target value and select the $r\%$ highest ranking features.

Modeling: Random Forest is used by constructing n decision trees

Hyperparameter Optimization: Gaussian Process

Experimental Results

Competitions

Peking University

KDD CUP 2014: Using past projects histories on DonrosChoose.org, predict if a crowd-funded project is exciting.

IJCAI: Using past merchant and customer shopping data, predict if a customer making a purchase during a promotional period will turn into a repeat buyer.

KDD CUP 2015: Using student interaction with resources on an online course, predict if the student will dropout in the next 10 days.

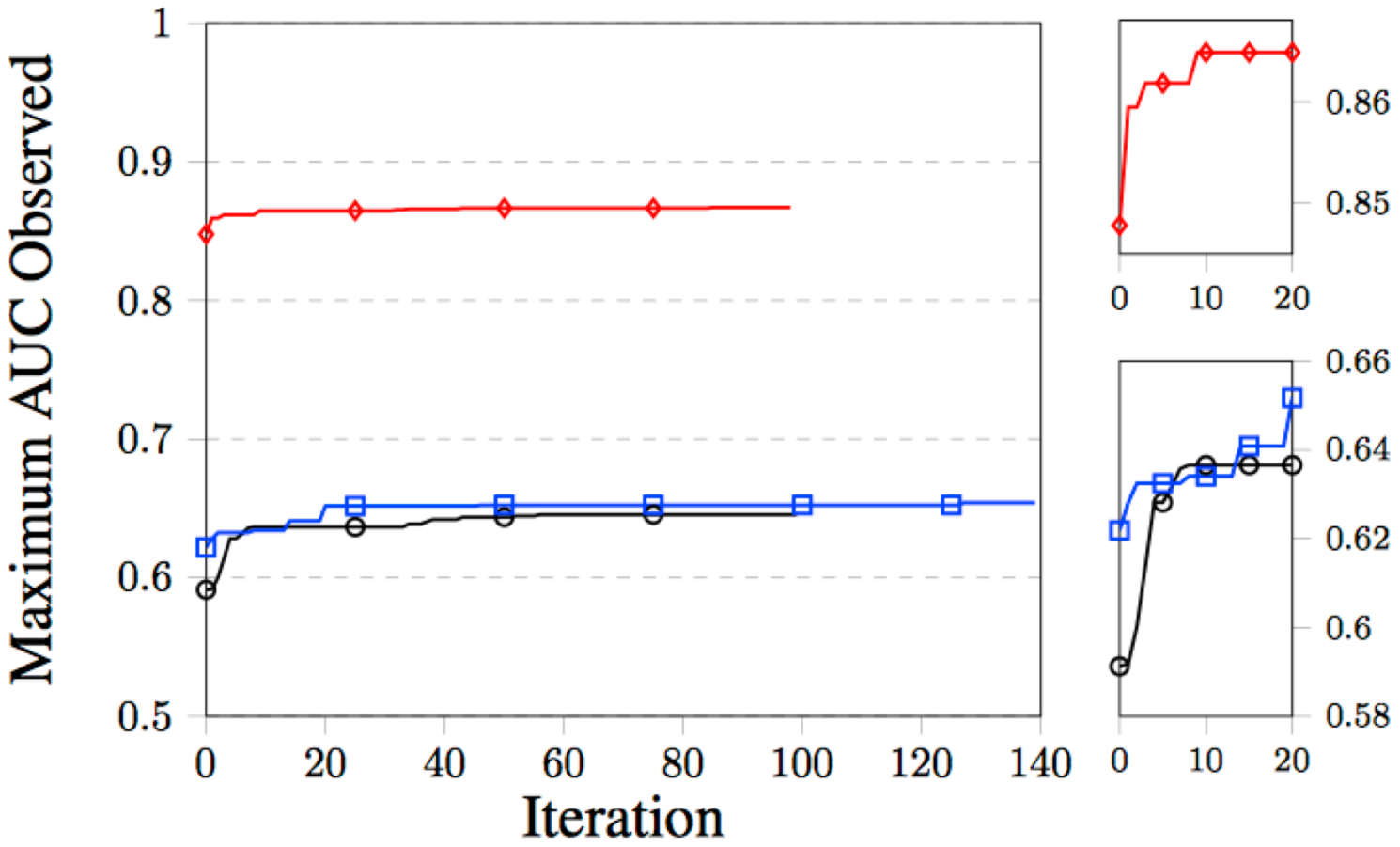
AUC Scores of Data Science Machine

Peking University

Parameter Selection	KDD cup 2014		IJCAI		KDD cup 2015	
	Local	Online	Local	Online	Local	Online
Default	.6059	.55481	.6439	.6313	.8444	.5000
GCP Optimized	.6321	.5863	.6540	.6606	.8672	.8621

The local score is the result of running 3-folds cross validation, while online score is based on submitting predictions to the competition.

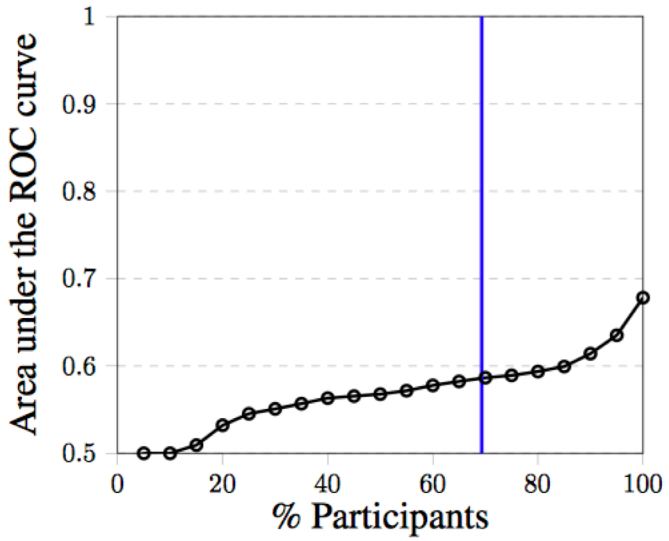
The Effect of Gaussian Optimization

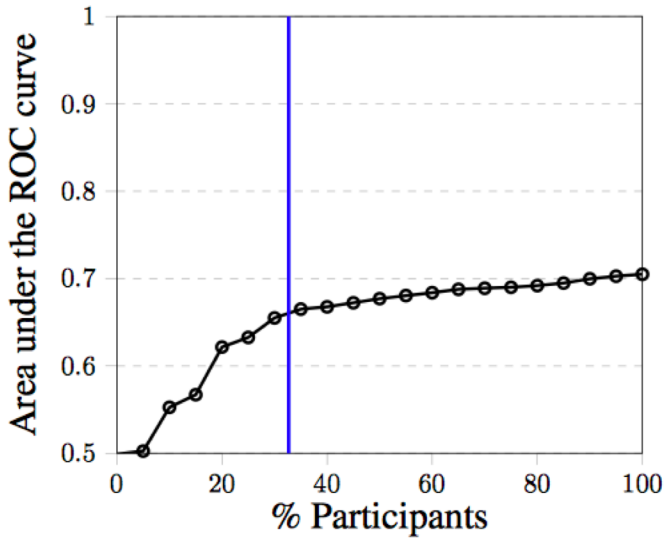


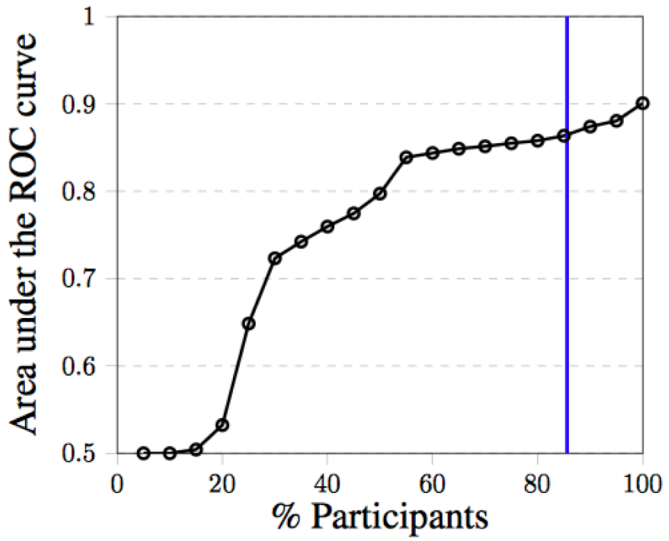
Performance Comparison

Peking University

Dataset	# Teams	% of Top Submission's Score	% of Teams Worse	# Submissions worse	# Days Spent on Worse Submissions
KDD cup 2014	473	86.5%	69.3%	3873	929
IJCAI	156	93.7%	32.7%	-	-
KDD cup 2015	277	95.7	85.6%	1319	377

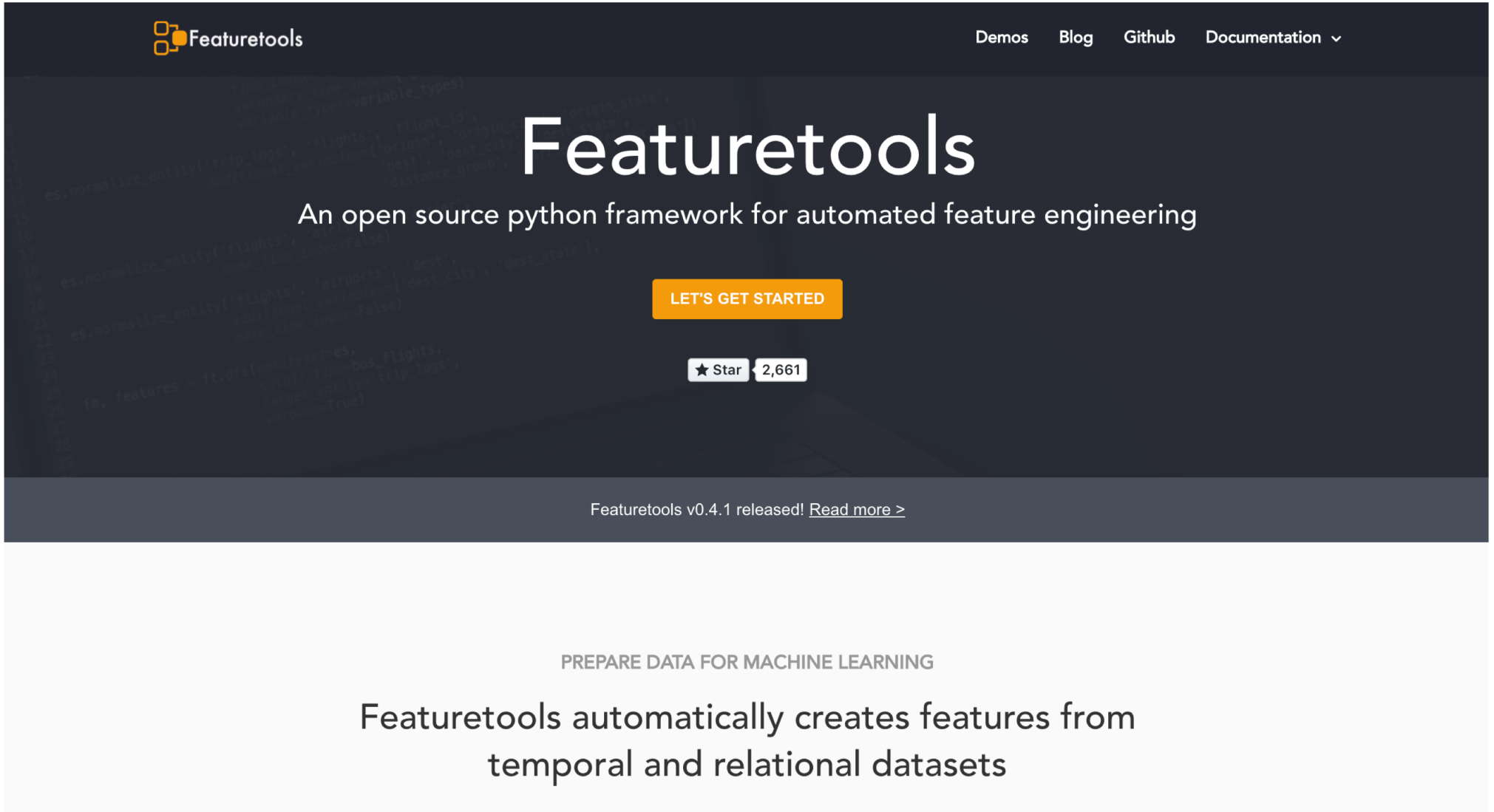






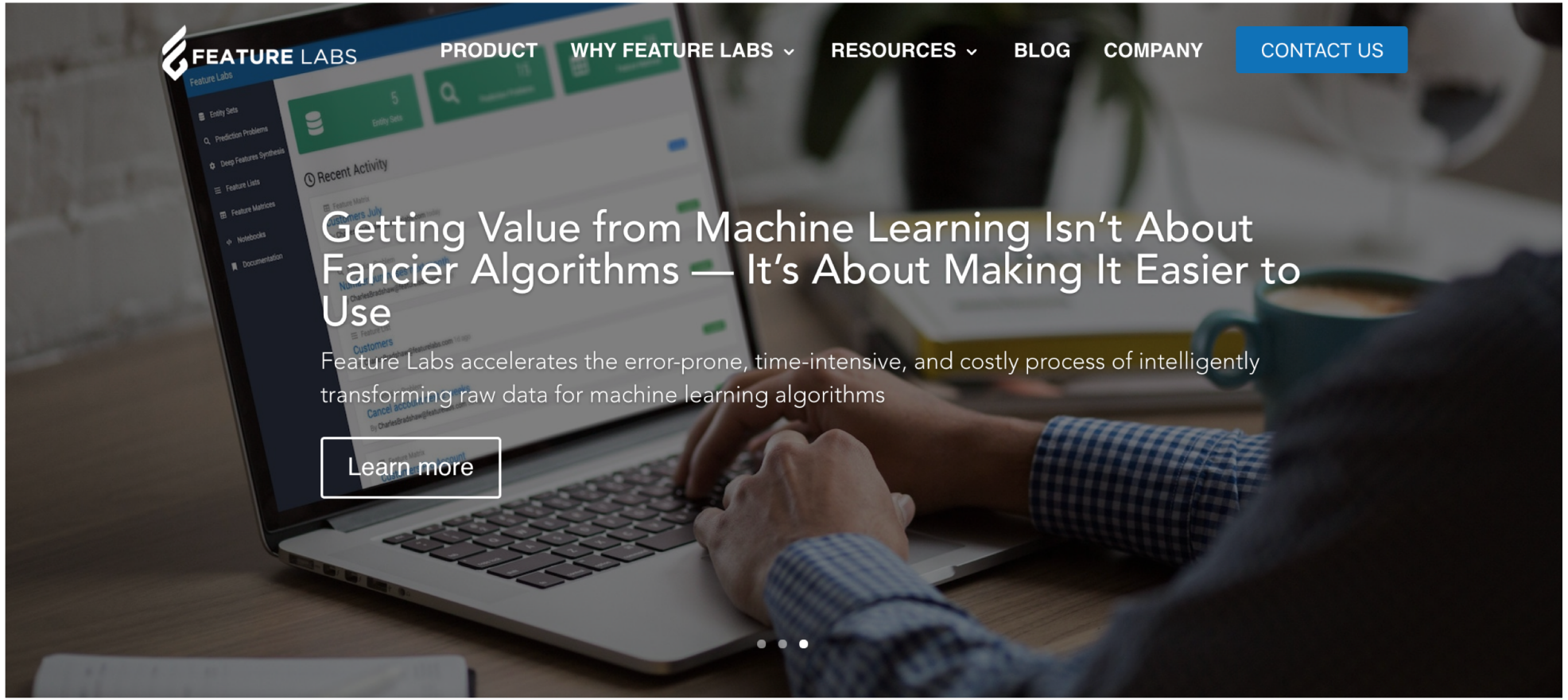
Featuretools

Peking University



Feature Labs

Peking University



Technology to remove the bottlenecks in the data science process

Conclusion

Peking University

Consider the scene with multi-tables which is suitable for industry.

The novelty is only focus on the part on auto feature engineering.

This method can be applied in the system in Tencent.

Peking University

Q & A