Meeting of the LF AI & Data Technical Advisory Council (TAC)

August 24, 2023

LF AI & DATA

Antitrust Policy

- Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.

- Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the TAC Wiki

Reminder: LF AI & Data Useful Links

- Web site: lfaidata.foundation
- Wiki: wiki.lfaidata.foundation
- GitHub: github.com/lfaidata
- Landscape: https://landscape.lfaidata.foundation or https://l.lfaidata.foundation
- Mail Lists: https://lists.lfaidata.foundation
- Slack: https://slack.lfaidata.foundation
- Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
- LF AI Logos: https://github.com/lfaidata/artwork/tree/master/lfaidata
- LF AI Presentation Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view?usp=sharing

- Events Page on LF AI Website: https://lfaidata.foundation/events/
- Events Calendar on LF AI Wiki (subscribe available): https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
- Event Wiki Pages: https://wiki.lfaidata.foundation/display/DL/LF+AI+Data+Foundation+Events

Agenda

- Roll Call (1 mins)
- Approval of Minutes from previous meeting (2 mins)
- SapientML – New project proposal
- Open Discussion

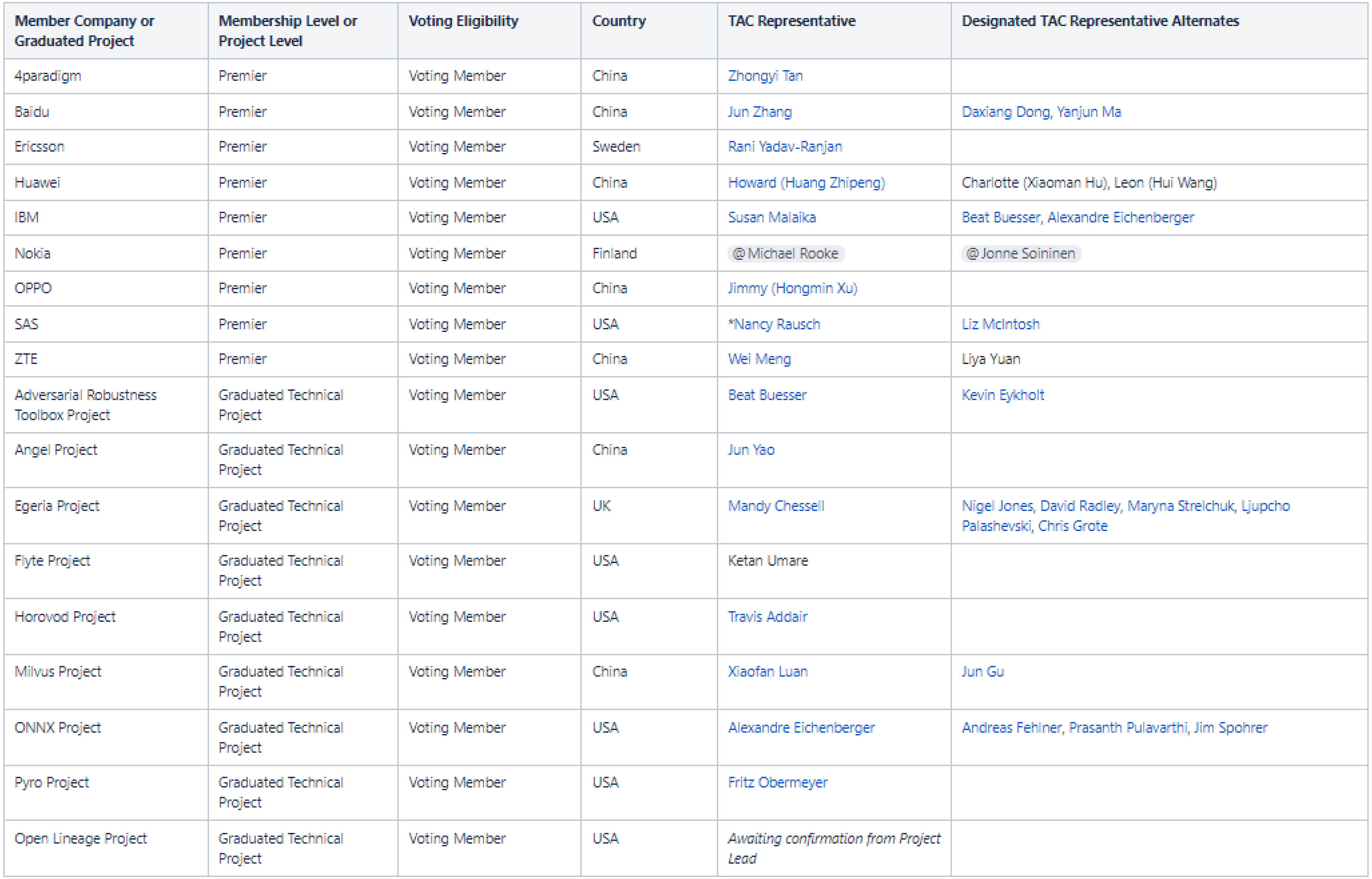
TAC Voting Members - Please note

Please ensure that you do the following to facilitate smooth procedural quorum and voting processes:

- Change your Zoom display name to include your First/Last Name, Company/Project Represented
  - example: Nancy Rausch, SAS
- State your First/Last Name and Company/Project when submitting a motion
  - example: First motion, Nancy Rausch/SAS

TAC Voting Members

Note: we still need a few designated backups specified on wiki



Member Company or Graduated Project	Membership Level or Project Level	Voting Eligibility	Country	TAC Representative	Designated TAC Representative Alternates
4paradigm	Premier	Voting Member	China	Zhongyi Tan	
Baidu	Premier	Voting Member	China	Jun Zhang	Daxiang Dong, Yanjun Ma
Ericsson	Premier	Voting Member	Sweden	Rani Yadav-Ranjan	
Huawei	Premier	Voting Member	China	Howard (Huang Zhipeng)	Charlotte (Xiaoman Hu), Leon (Hui Wang)
IBM	Premier	Voting Member	USA	Susan Malaika	Beat Buesser, Alexandre Eichenberger
Nokia	Premier	Voting Member	Finland	@Michael Rooke	@Jonne Soininen
OPPO	Premier	Voting Member	China	Jimmy (Hongmin Xu)	
SAS	Premier	Voting Member	USA	*Nancy Rausch	Liz McIntosh
ZTE	Premier	Voting Member	China	Wei Meng	Liya Yuan
Adversarial Robustness Toolbox Project	Graduated Technical Project	Voting Member	USA	Beat Buesser	Kevin Eykholt
Angel Project	Graduated Technical Project	Voting Member	China	Jun Yao	
Egeria Project	Graduated Technical Project	Voting Member	UK	Mandy Chessell	Nigel Jones, David Radley, Maryna Strelchuk, Ljupcho Palashevski, Chris Grote
Flyte Project	Graduated Technical Project	Voting Member	USA	Ketan Umare	
Horovod Project	Graduated Technical Project	Voting Member	USA	Travis Addair	
Milvus Project	Graduated Technical Project	Voting Member	China	Xiaofan Luan	Jun Gu
ONNX Project	Graduated Technical Project	Voting Member	USA	Alexandre Eichenberger	Andreas Fehlner, Prasanth Pulavarthi, Jim Spohrer
Pyro Project	Graduated Technical Project	Voting Member	USA	Fritz Obermeyer	
Open Lineage Project	Graduated Technical Project	Voting Member	USA	*Awaiting confirmation from Project Lead*	

Minutes approval

Approval of August 10, 2023 Minutes

Draft minutes from the August 10 TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

- That the minutes of the August 10 meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

LF AI & DATA

Fujitsu's Proposal to Host SapientML project in LF AI & Data

Hiro Kobashi (hkobashi@fujitsu.com)

Masahiro Fukuyori (fukuyori@fujitsu.com)

(Representing AutoML team at Fujitsu)

2023-08-24

Why contribute SapientML to Linux Foundation

- Neutral holding ground
  - vendor-neutral, not for profit
- Open governance model
  - Transparent and open governance model
  - Instill trust in contributors and adopters in the management of the project
  - Neutral management of projects’ assets by the foundation
- Growing community
  - Increase visibility of project through LF ecosystem
  - Increase contributors by converting new & existing users
  - Opportunities to collaborate with other projects

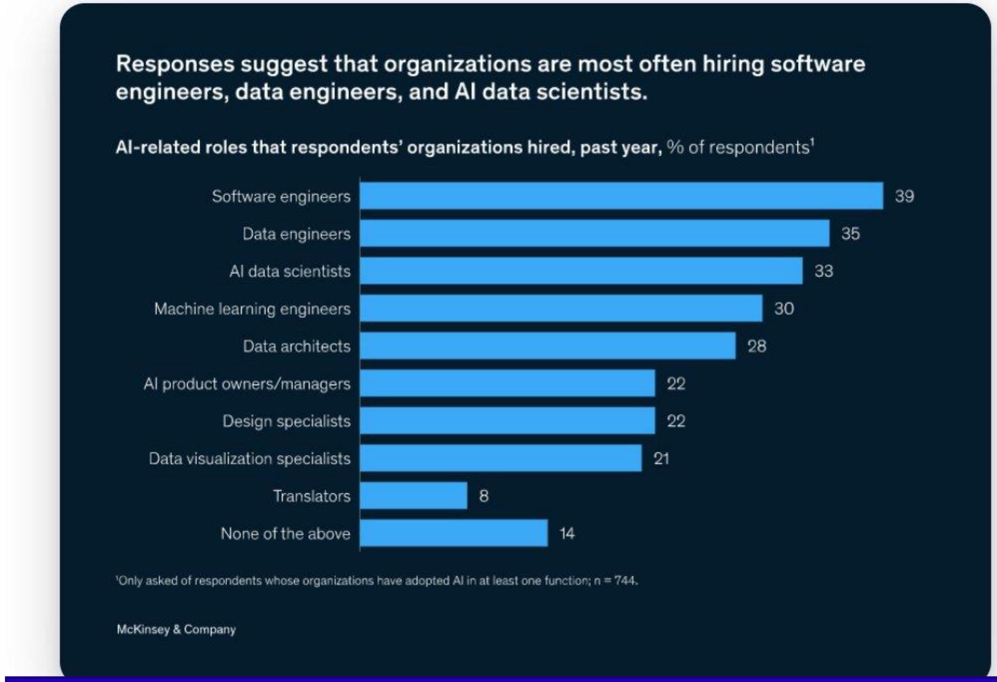
Agenda

- 1. Background
- 2. Challenges
- 3. How SapientML helps
- 4. Next steps

Background

The shortage of AI talent has not been resolved.

It is still difficult to recruit AI talent.



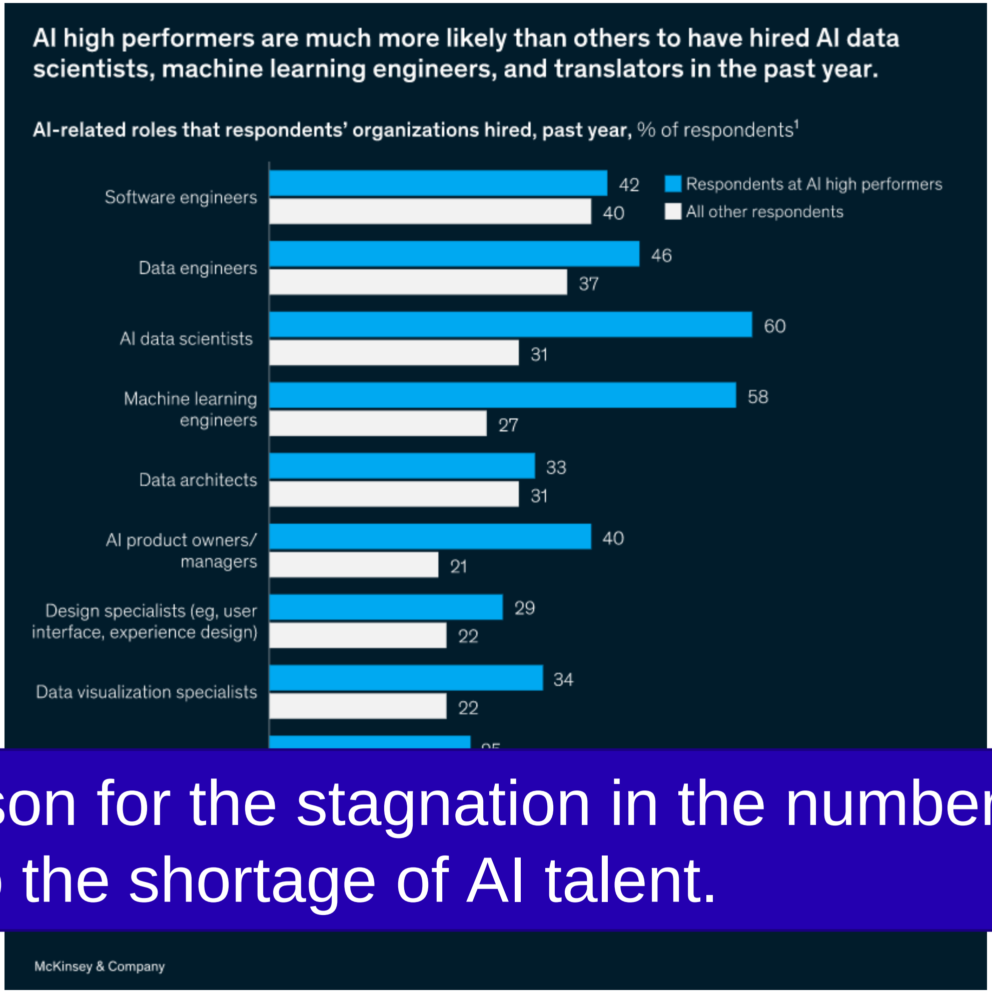
Responses suggest that organizations are most often hiring software engineers, data engineers, and AI data scientists.

AI-related roles that respondents' organizations hired, past year, % of respondents[1]

[1] Only asked of respondents whose organizations have adopted AI in at least one function; n = 744.

McKinsey & Company

In particular, there is a shortage of AI talent such as data scientists and data engineers.



AI high performers are much more likely than others to have hired AI data scientists, machine learning engineers, and translators in the past year.

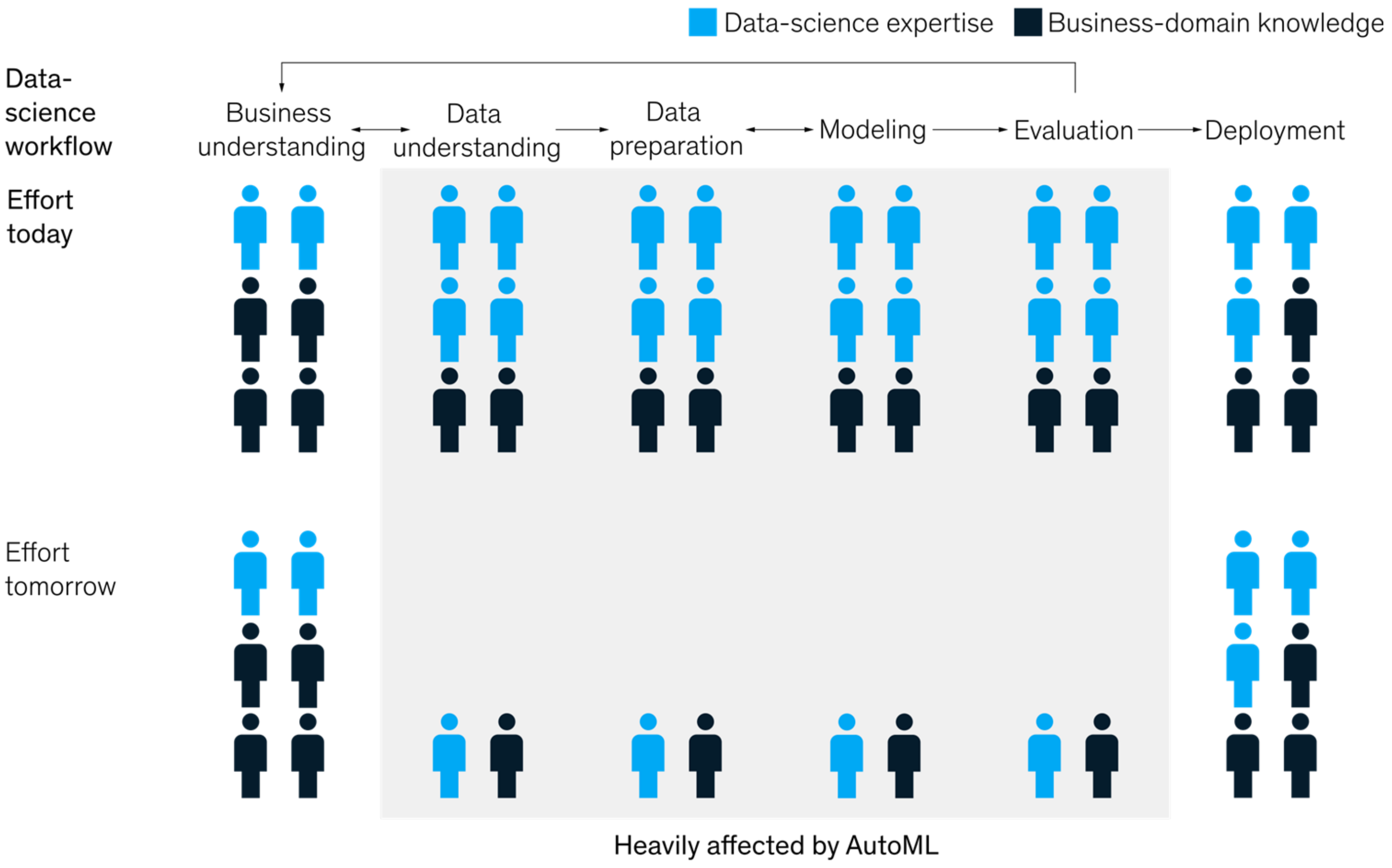
AI-related roles that respondents' organizations hired, past year, % of respondents[1]

McKinsey & Company

It is my personal speculation that the reason for the stagnation in the number of AI application cases is due to the shortage of AI talent.

The state of AI in 2022—and a half decade in review by McKinsey

AutoML

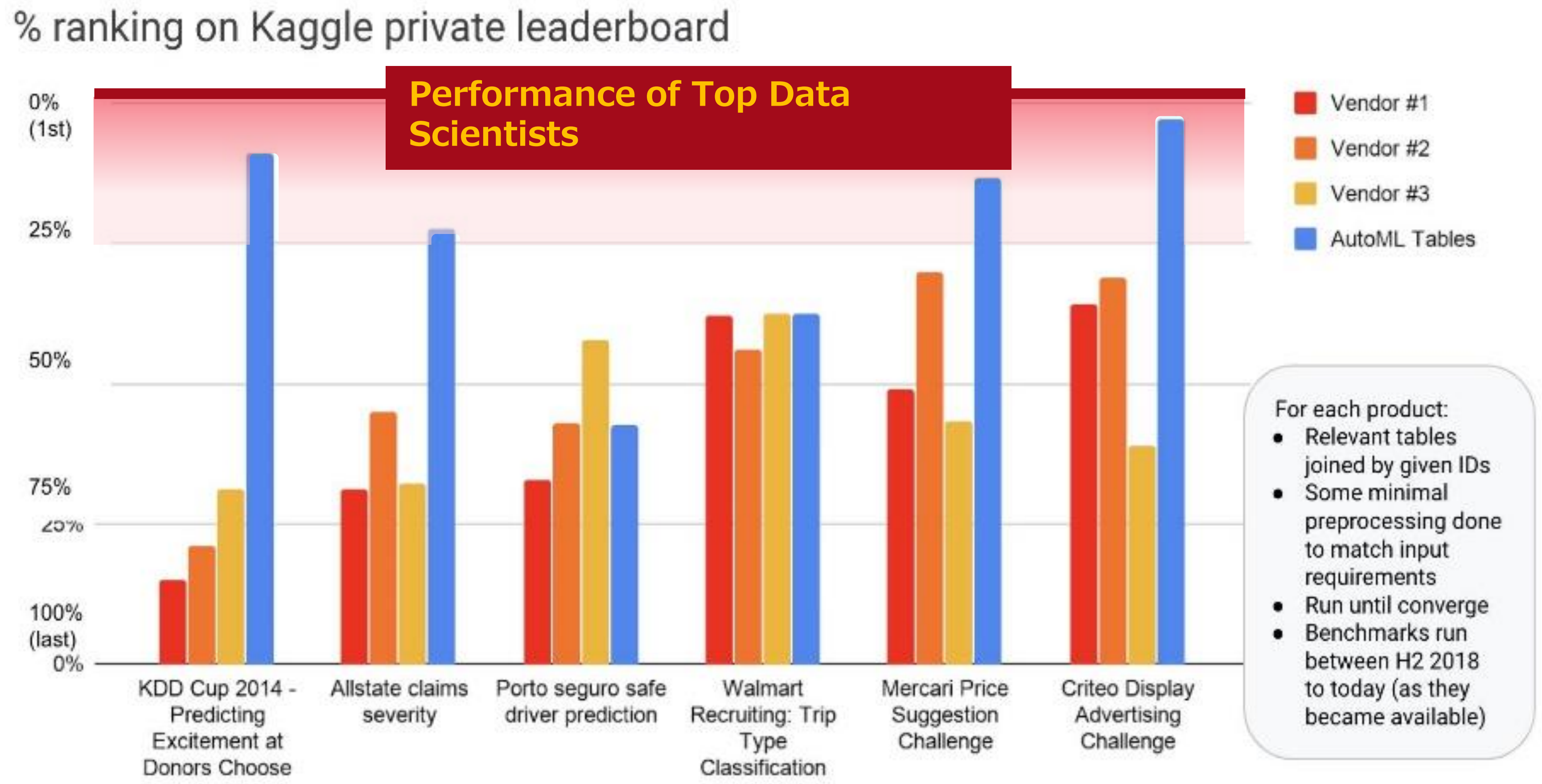


AutoML is a research activity that **scales AI application by automating AI utilization processes**.

https://www.mckinsey.com/capabilities/quantumblack/our-insights/rethinking-ai-talent-strategy-as-automated-machine-learning-comes-of-age

AutoML: Reality or Pipe Dream?

AutoML beginning to rival top data scientists in ***some*** cases



Performance of Top Data Scientists

For each product:

- Relevant tables joined by given IDs
- Some minimal preprocessing done to match input requirements
- Run until converge
- Benchmarks run between H2 2018 to today (as they became available)

Source: Google Cloud Next 2019 talk

AutoML: Reality or Pipe Dream?

FUJITSU

AI model creation is rapidly automated (AutoML)

Issue in AI model creation

○ Many try-and-error are needed (taking long time)

○ Model quality is highly depending on individual skill

AutoML evolution

- The emergence of AutoML, that performs the same level of top data scientist

California Design Den

Online sale of bedding (*1)

Quickly create AI model with AutoML without data scientists

Reduce 50% of stocked items in factories

G5

Real Estate Customer Acquisition Business (*2)

AI model generated by AutoML is as accurate as its by top data scientists

Reduce 80% time and record 95% accuracy

Lenovo

Device sales (*3)

AutoML enables rapid deployment of accurate models

Reduced model build time from 4 weeks to 3 days

Sources: *1: https://cloud.google.com/customers/california-design-den,
*2: https://h2o.ai/case-studies/driving-marketing-performance-with-machine-learning/,
*3: https://aws.amazon.com/jp/partners/success/lenovo-datarobot/

Challenges

Challenges

- AutoML tools have not yet been adopted widely. Why?

1. **Time-consuming**/Needs intensive computational resources
2. **Blackbox nature** of the generated AI models

Key factors to success

1. **Speed (in AI model creation)**
   - How quickly can we create an AI model to fit in the data scientist's routine
2. **Code (with explanation)**
   - Code synthesis with explanation can enable iterations (try&error), which are the matter in AI process.
3. **Accuracy**
   - No doubt, accuracy is always matter in AI model creation.

SapientML

Meta-Learning based AutoML to bring successes to AI model creation process.

Approach

1 Extract expert knowledge from codes

Break the limit (1)
Current AutoML searches parameters from pre-defined search space. We utilize expert experience, which are stored in codes.

2 Be interactive with data scientist

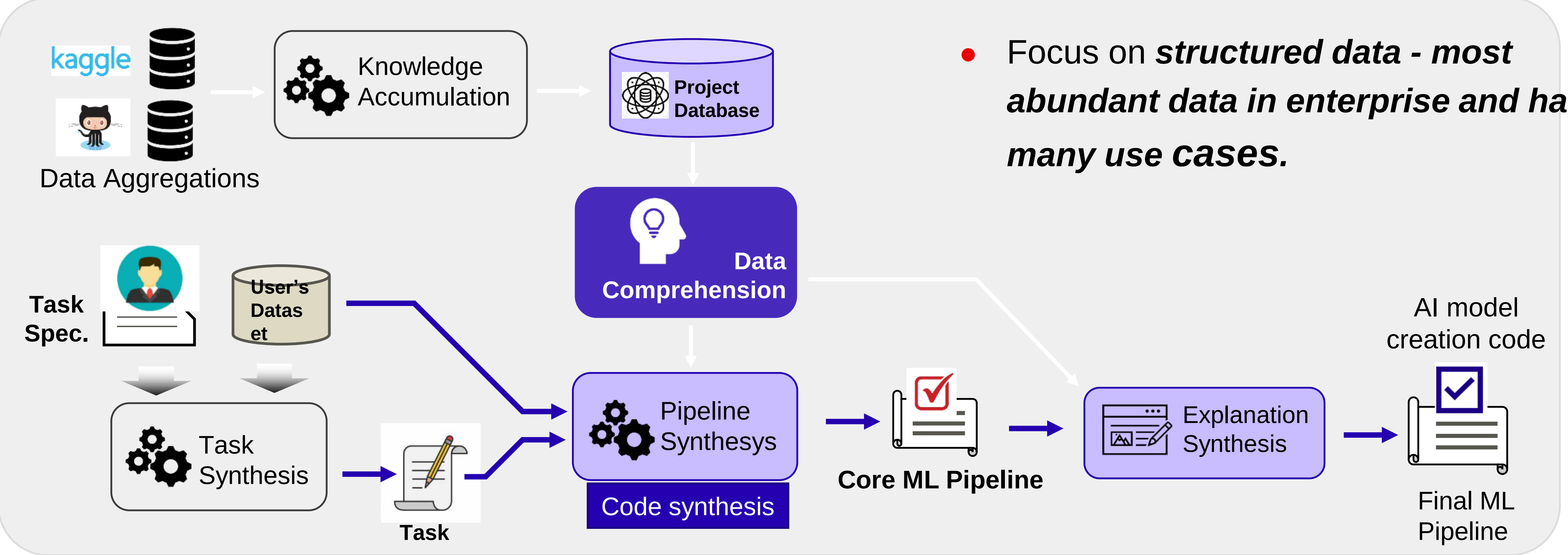
Break the limit (2)
Current AutoML creates AI model only. There is no flexibility (ex. No chance to modify, hard to understand why the model is good)

Meta-Learning based AutoML

- Synthesis high accurate **AI model creation code** by using expert knowledge
- Provide the modifiable code to interact with data scientist

Generative AutoML: SapientML

- AutoML that harnesses expertise of DS encapsulated in large corpora of existing ML pipelines, e.g., Kaggle



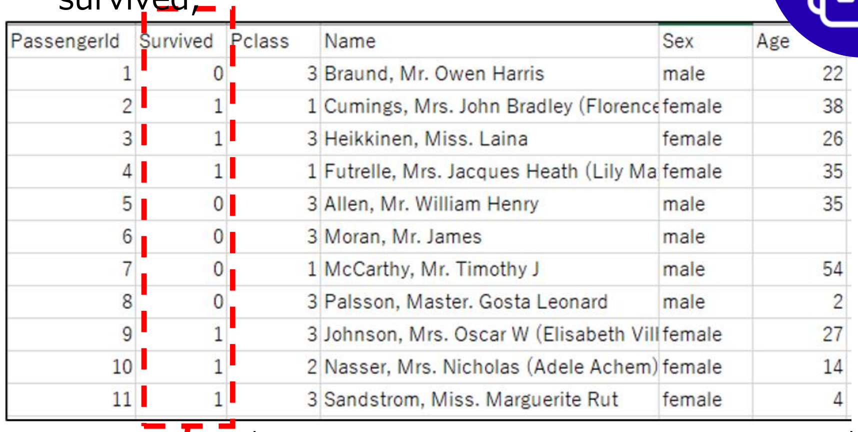
- Focus on ***structured data - most abundant data in enterprise and has many use cases.***

SapientML: An AutoML approach harnessing the wisdom (sapere) of human (sapien) data scientists.

How works SapientML?

- Input Table Data and ML Spec. ☐ Output AI model with code

e.g. When you build an AI model that predicts whether or not each Titanic passenger survived,



PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath (Lily Ma	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Moran, Mr. James	male	
7	0	1	McCarthy, Mr. Timothy J	male	54
8	0	3	Palsson, Master. Gosta Leonard	male	2
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vil	female	27
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4

[Target Variable]
"survived?"
(Prediction Target)

[Explanatory Variable]
Various attributes of passengers

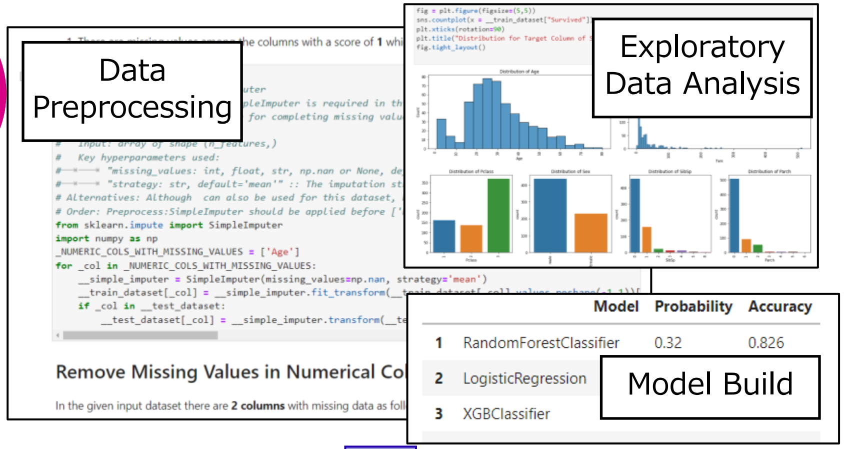
Table Data

Input

Output

SapientML

Machine Learning Program



Data Preprocessing

1. There are missing values among the columns with a score of 1 whi

```
#         uter
#   pleImputer is required in th
#   for completing missing valu
#   Input: array of shape (n_features,)
#   Key hyperparameters used:
#→ ※ → "missing_values: int, float, str, np.nan or None, de
#→ ※ → "strategy: str, default='mean'" :: The imputation st
# Alternatives: Although  can also be used for this dataset,
# Order: Preprocess:SimpleImputer should be applied before ['
from sklearn.impute import SimpleImputer
import numpy as np
_NUMERIC_COLS_WITH_MISSING_VALUES = ['Age']
for _col in _NUMERIC_COLS_WITH_MISSING_VALUES:
    __simple_imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
    __train_dataset[_col] = __simple_imputer.fit_transform(__train_dataset[_col].values.reshape(-1,1))[
    if _col in __test_dataset:
        __test_dataset[_col] = __simple_imputer.transform(__te
```

Remove Missing Values in Numerical Col

In the given input dataset there are **2 columns** with missing data as foll

Exploratory Data Analysis

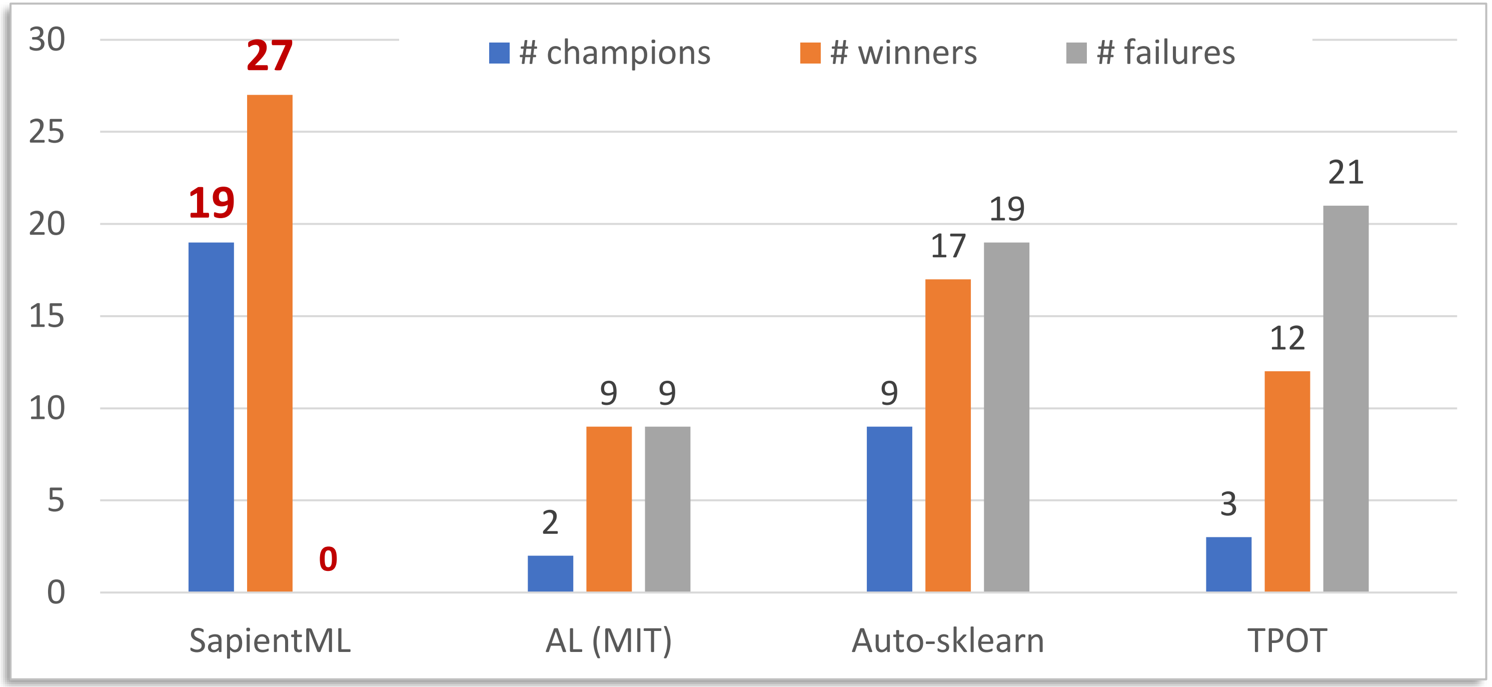
```
fig = plt.figure(figsize=(5,5))
sns.countplot(x = __train_dataset["Survived"])
plt.xticks(rotation=90)
plt.title("Distribution for Target Column of S
fig.tight_layout()
```

Model Build

	Model	Probability	Accuracy
1	RandomForestClassifier	0.32	0.826
2	LogisticRegression		
3	XGBClassifier		

Execute output program

SapientML at ICSE2022



SAPIENTML: Synthesizing Machine Learning Pipelines by Learning from Human-Written Solutions
Ripon Saha, Akira Ura, Sonal Mahajan, Chenguang Zhu, Linyi Li, Yang Hu, Hiroaki Yoshida, Sarfraz Khurshid, Mukul Prasad

May 21-29, 2022
Pittsburgh, PA,
USA

S-rank, CORE: A*
Flagship SE Conference

Acceptance rate:
26% (197/751)

- SapientML out-performs all academic AutoML tools on incl. SoTA AL (MIT)
  - 41 benchmarks, including 20 Kaggle competitions
- No failures, highest number of wins
- SapientML synthesis is quite robust to variations in training data – pipelines should generalize well

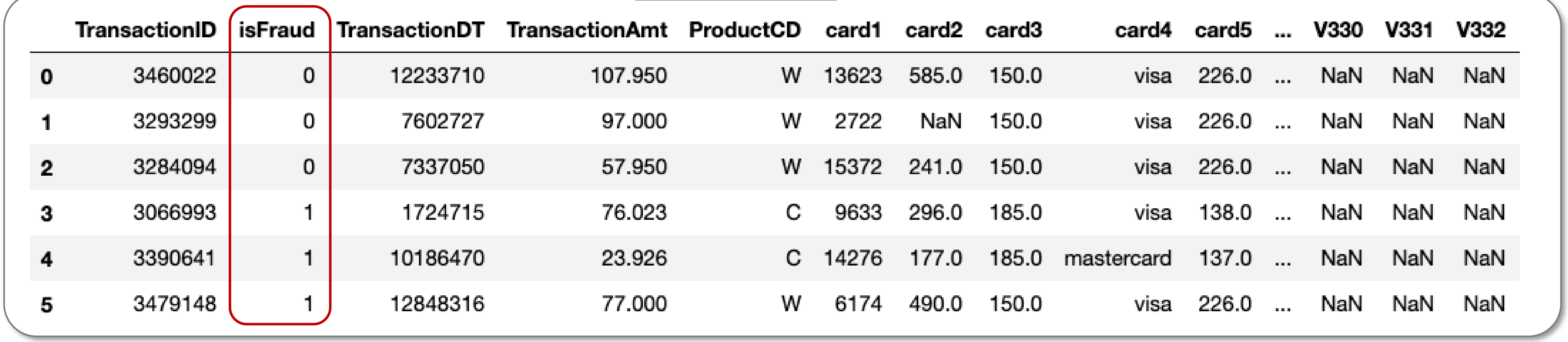
Illustrative Example

FUJITSU

Example: IEEE-CIS-Fraud-Detection (Kaggle Competition)

Rows: e-commerce transactions, represented by 394 features, device type, product features, etc.

Dataset



	TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4	card5	...	V330	V331	V332
0	3460022	0	12233710	107.950	W	13623	585.0	150.0	visa	226.0	...	NaN	NaN	NaN
1	3293299	0	7602727	97.000	W	2722	NaN	150.0	visa	226.0	...	NaN	NaN	NaN
2	3284094	0	7337050	57.950	W	15372	241.0	150.0	visa	226.0	...	NaN	NaN	NaN
3	3066993	1	1724715	76.023	C	9633	296.0	185.0	visa	138.0	...	NaN	NaN	NaN
4	3390641	1	10186470	23.926	C	14276	177.0	185.0	mastercard	137.0	...	NaN	NaN	NaN
5	3479148	1	12848316	77.000	W	6174	490.0	150.0	visa	226.0	...	NaN	NaN	NaN

Task: Classifying if a transaction is ***Fraud***

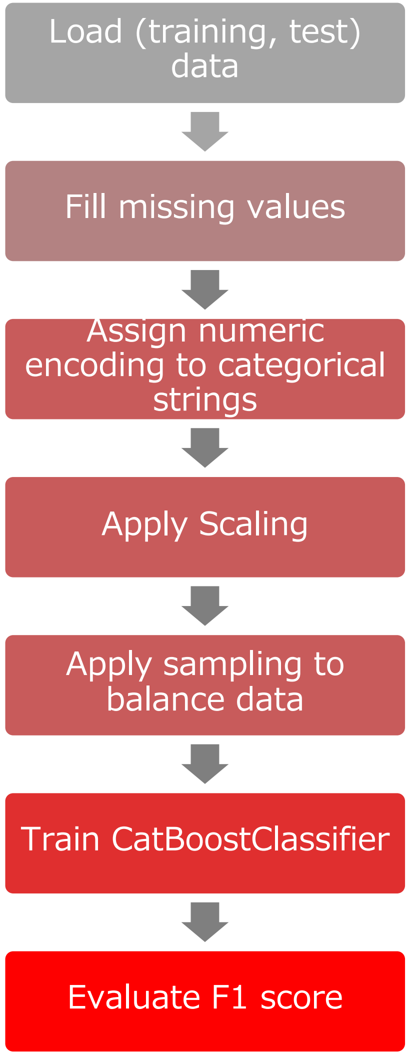
→ **Problem type:** (binary) classification

→ **Target column:** **isFraud** (*1 or 0*)

SpaientML generated pipeline

# shortened for presentation

```
# LOAD DATA
__train_dataset=pd.read_csv("training.csv", delimiter=",")
__test_dataset=pd.read_csv("test.csv", delimiter=",")
# PREPROCESSING-1
_STRING_CATG_COLM_HAS_MISSING = ['card4', 'card6', 'P_emaildomain’,…]
for _col in _STRING_COLS_WITH_MISSING_VALUES:
   __si = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
   __train_dataset[_col] = __si.fit_transform(__train_dataset[_col].values.reshape(-1,1))[:,0]
   __test_dataset[_col] = _si.transform(__test_dataset[_col].astype(__train_dataset[_col].dtypes).values.reshape(-1,1))[:,0]
# PREPROCESSING-2
_CAT_COLS = ['ProductCD', 'card4', 'card6’, …,'M9']
_ohe = OrdinalEncoder(handle_unknown="use_encoded_value", unknown_value=-1)
__train_dataset[_CAT_COLS] = pd.DataFrame(_ohe.fit_transform(__train_dataset[_CAT_COLS]), columns=_CAT_COLS)
__test_dataset[_CAT_COLS] = pd.DataFrame(_ohe.transform(__test_dataset[_CAT_COLS]), columns=_CAT_COLS)
# PREPROCESSING-4
from sklearn.preprocessing import StandardScaler
__ss= StandardScaler()
__feature_train = pd.DataFrame(__ss.fit_transform(__feature_train.values), index=__feature_train.index, columns=__feature_train.columns)
__feature_test = pd.DataFrame(__ss.transform(__feature_test.values), index=__feature_test.index, columns=__feature_test.columns)
# PREPROCESSING-5
from imblearn.over_sampling import SMOTE
smote = SMOTE()
__feature_train, __target_train = smote.fit_resample(__feature_train, __target_train)
# MODEL
from catboost import CatBoostClassifier
__model = CatBoostClassifier()
__model.fit(__feature_train, __target_train)
__y_pred = __model.predict(__feature_test)
# EVALUATION
from sklearn import metrics
__f1 = metrics.f1_score(__target_test, __y_pred, average='macro')
print('RESULT: F1 Score: ' + str(__f1))
```



Load (training, test) data

Fill missing values

Assign numeric encoding to categorical strings

Apply Scaling

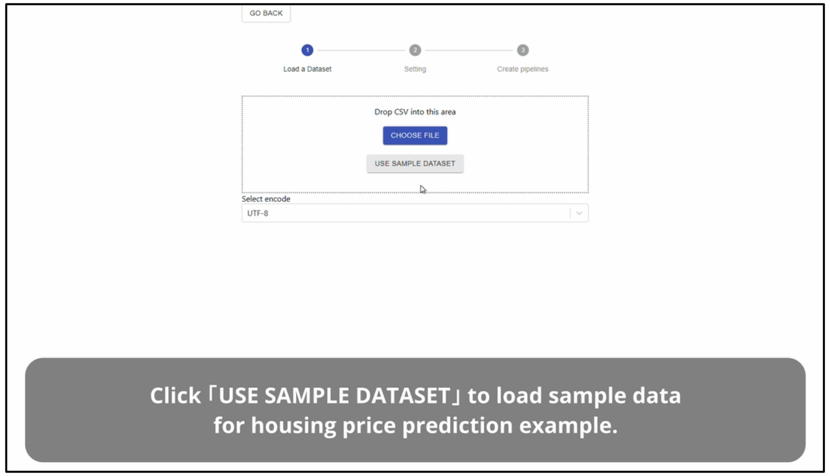
Apply sampling to balance data

Train CatBoostClassifier

Evaluate F1 score

GUI – Simple 3 steps to generate AI

1.Load dataset

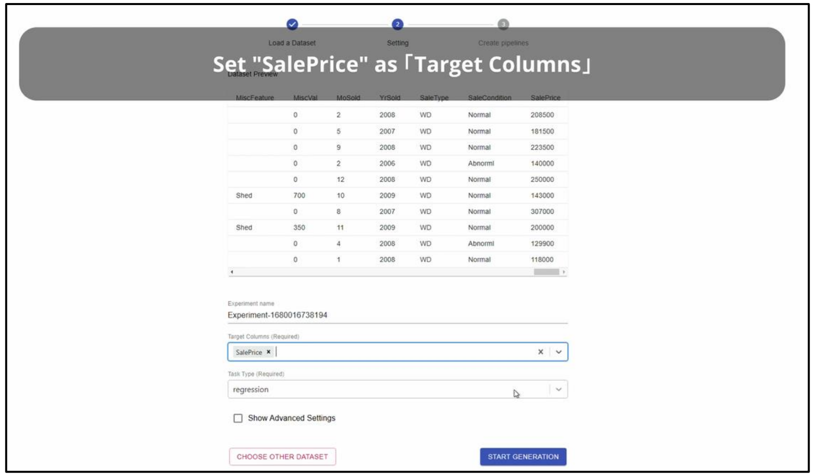


Drop CSV into this area

Select encode: UTF-8

Click「USE SAMPLE DATASET」to load sample data for housing price prediction example.

2.Specify ML task



Set "SalePrice" as「Target Columns」

Dataset Preview

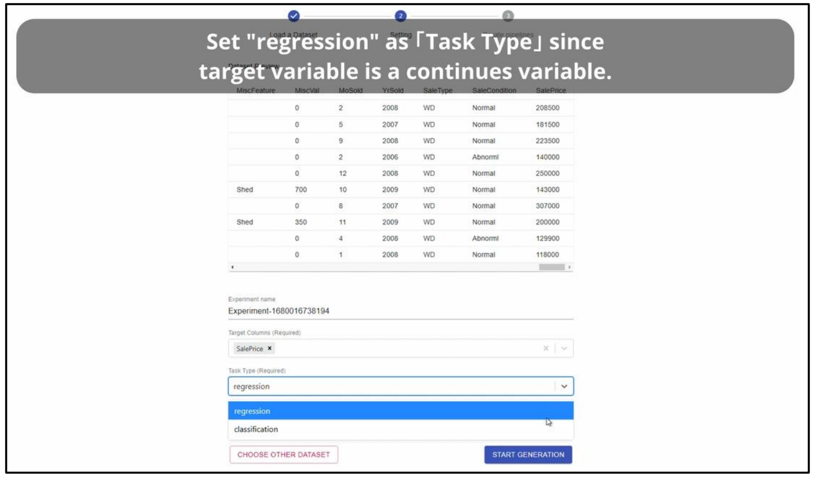
MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
	0	2	2008	WD	Normal	208500
	0	5	2007	WD	Normal	181500
	0	9	2008	WD	Normal	223500
	0	2	2006	WD	Abnorml	140000
	0	12	2008	WD	Normal	250000
Shed	700	10	2009	WD	Normal	143000
	0	8	2007	WD	Normal	307000
Shed	350	11	2009	WD	Normal	200000
	0	4	2008	WD	Abnorml	129900
	0	1	2008	WD	Normal	118000

Experiment name: Experiment-1680016738194

Target Columns (Required): SalePrice ×

Task Type (Required): regression

□ Show Advanced Settings



Set "regression" as「Task Type」since target variable is a continues variable.

MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
	0	2	2008	WD	Normal	208500
	0	5	2007	WD	Normal	181500
	0	9	2008	WD	Normal	223500
	0	2	2006	WD	Abnorml	140000
	0	12	2008	WD	Normal	250000
Shed	700	10	2009	WD	Normal	143000
	0	8	2007	WD	Normal	307000
Shed	350	11	2009	WD	Normal	200000
	0	4	2008	WD	Abnorml	129900
	0	1	2008	WD	Normal	118000

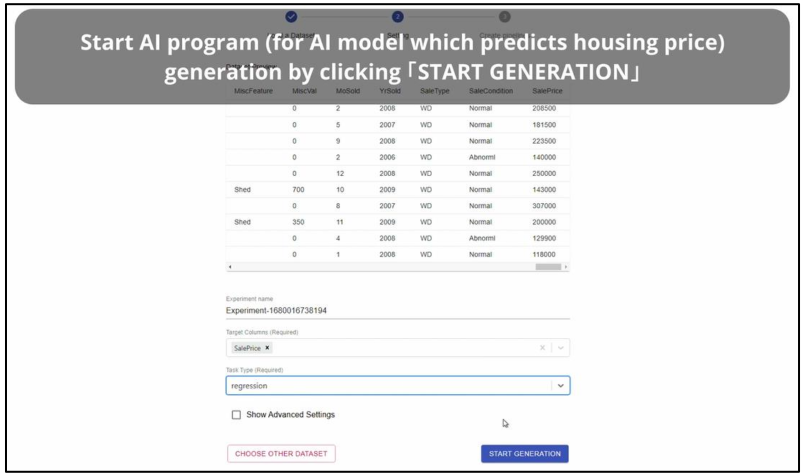
Experiment name: Experiment-1680016738194

Target Columns (Required): SalePrice ×

Task Type (Required): regression

- regression
- classification

3. Request to generate



Start AI program (for AI model which predicts housing price) generation by clicking「START GENERATION」

MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
	0	2	2008	WD	Normal	208500
	0	5	2007	WD	Normal	181500
	0	9	2008	WD	Normal	223500
	0	2	2006	WD	Abnorml	140000
	0	12	2008	WD	Normal	250000
Shed	700	10	2009	WD	Normal	143000
	0	8	2007	WD	Normal	307000
Shed	350	11	2009	WD	Normal	200000
	0	4	2008	WD	Abnorml	129900
	0	1	2008	WD	Normal	118000

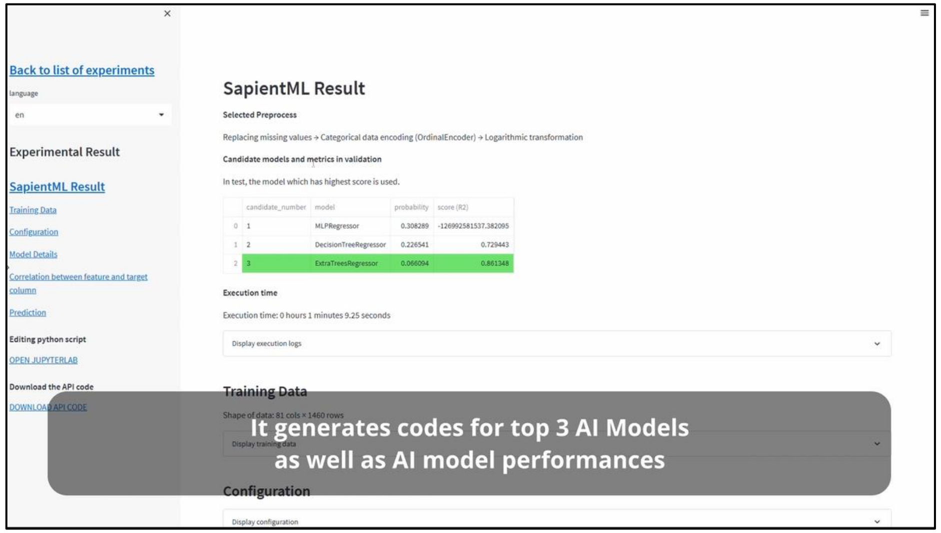
Experiment name: Experiment-1680016738194

Target Columns (Required): SalePrice ×

Task Type (Required): regression

□ Show Advanced Settings

GUI – Prediction using generated AI



SapientML Result

Selected Preprocess

Replacing missing values → Categorical data encoding (OrdinalEncoder) → Logarithmic transformation

Candidate models and metrics in validation

In test, the model which has highest score is used.

	candidate_number	model	probability	score (R2)
0	1	MLPRegressor	0.308289	-126992581537.382095
1	2	DecisionTreeRegressor	0.226541	0.729443
2	3	ExtraTreesRegressor	0.066094	0.861348

Execution time

Execution time: 0 hours 1 minutes 9.25 seconds

Display execution logs

Training Data

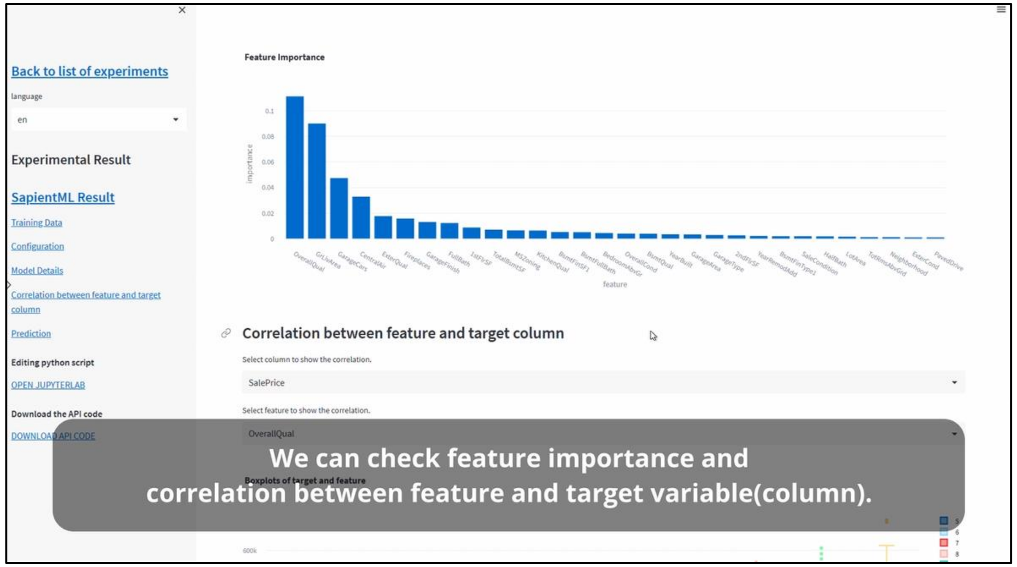
Shape of data: 81 cols × 1460 rows

Display training data

Configuration

Display configuration

It generates codes for top 3 AI Models
as well as AI model performances



Correlation between feature and target column

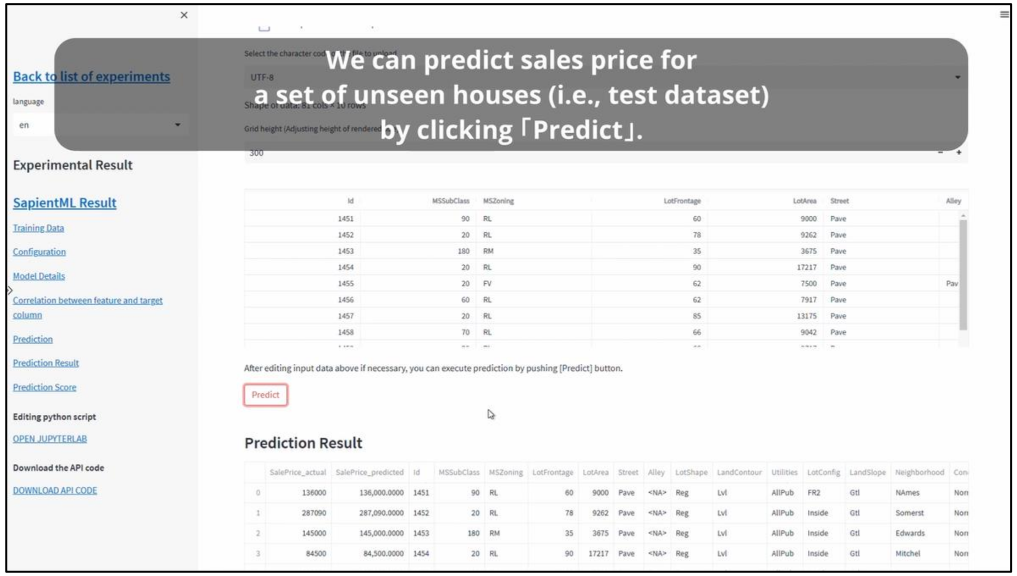
Select column to show the correlation.

SalePrice

Select feature to show the correlation.

OverallQual

We can check feature importance and
correlation between feature and target variable(column).



Select the character code of the file to upload

UTF-8

Shape of data: 81 cols × 10 rows

Grid height (Adjusting height of rendered

300

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley
1451	90	RL	60	9000	Pave	
1452	20	RL	78	9262	Pave	
1453	180	RM	35	3675	Pave	
1454	20	RL	90	17217	Pave	
1455	20	FV	62	7500	Pave	Pav
1456	60	RL	62	7917	Pave	
1457	20	RL	85	13175	Pave	
1458	70	RL	66	9042	Pave	

After editing input data above if necessary, you can execute prediction by pushing [Predict] button.

Predict

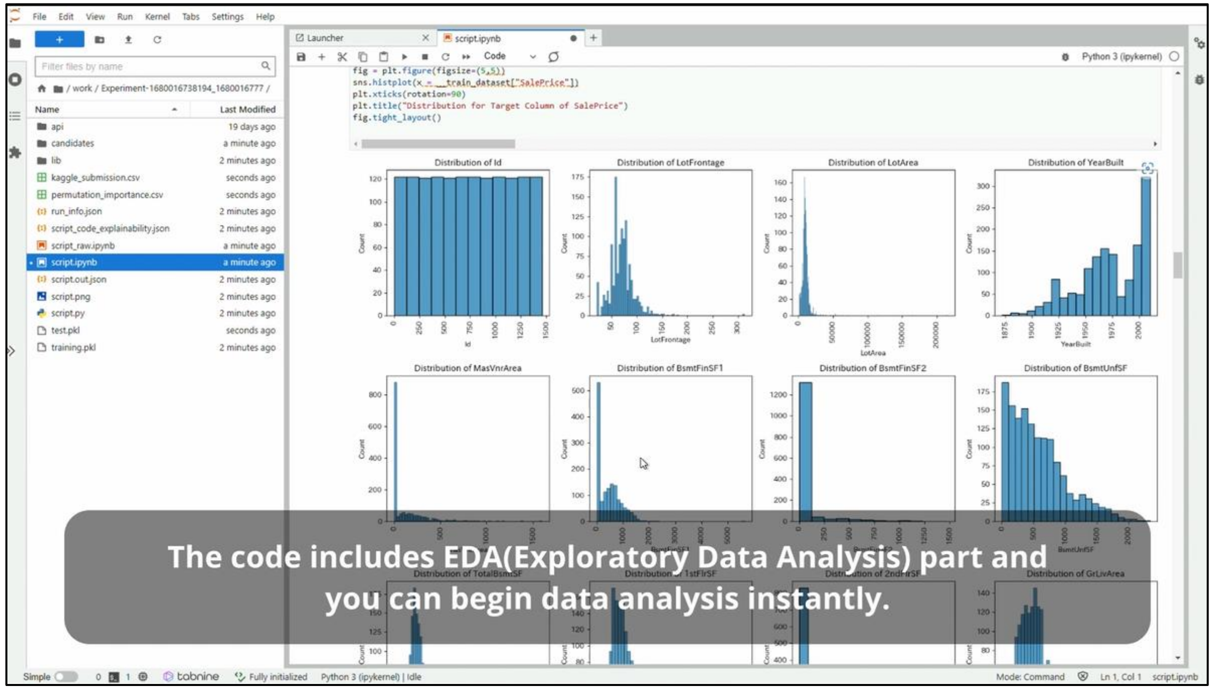
Prediction Result

	SalePrice_actual	SalePrice_predicted	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Con
0	136000	136,000.0000	1451	90	RL	60	9000	Pave	<NA>	Reg	Lvl	AllPub	FR2	Gtl	NAmes	Nor
1	287090	287,090.0000	1452	20	RL	78	9262	Pave	<NA>	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Nor
2	145000	145,000.0000	1453	180	RM	35	3675	Pave	<NA>	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Nor
3	84500	84,500.0000	1454	20	RL	90	17217	Pave	<NA>	Reg	Lvl	AllPub	Inside	Gtl	Mitchel	Nor

We can predict sales price for
a set of unseen houses (i.e., test dataset)
by clicking「Predict」.

GUI – EDA & Code of generated AI

Start Exploratory Data Analysis instantly



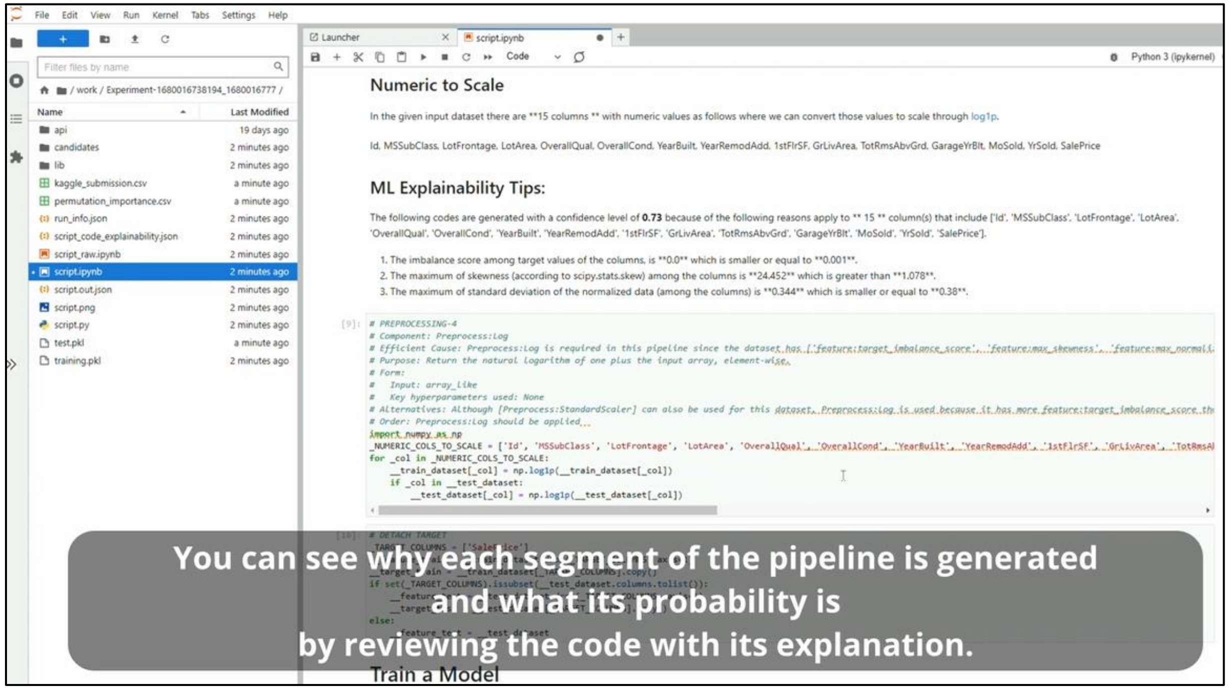
File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ work / Experiment-1680016738194_1680016777 /
Launcher
script.ipynb
Code
Python 3 (ipykernel)
Distribution of Id
Distribution of LotFrontage
Distribution of LotArea
Distribution of YearBuilt
Distribution of MasVnrArea
Distribution of BsmtFinSF1
Distribution of BsmtFinSF2
Distribution of BsmtUnfSF
Distribution of TotalBsmtSF
Distribution of 1stFlrSF
Distribution of 2ndFlrSF
Distribution of GrLivArea
Count
Simple
tabnine
Fully initialized
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
script.ipynb

Name	Last Modified
api	19 days ago
candidates	a minute ago
lib	2 minutes ago
kaggle_submission.csv	seconds ago
permutation_importance.csv	seconds ago
run_info.json	2 minutes ago
script_code_explainability.json	2 minutes ago
script_raw.ipynb	a minute ago
script.ipynb	a minute ago
script.out.json	2 minutes ago
script.png	2 minutes ago
script.py	2 minutes ago
test.pkl	seconds ago
training.pkl	2 minutes ago

```
fig = plt.figure(figsize=(5,5))
sns.histplot(x = __train_dataset["SalePrice"])
plt.xticks(rotation=90)
plt.title("Distribution for Target Column of SalePrice")
fig.tight_layout()
```

The code includes EDA(Exploratory Data Analysis) part and you can begin data analysis instantly.

Customize AI model easily by modifying generated code



File Edit View Run Kernel Tabs Settings Help
Filter files by name
/ work / Experiment-1680016738194_1680016777 /
Launcher
script.ipynb
Code
Python 3 (ipykernel)

Name	Last Modified
api	19 days ago
candidates	2 minutes ago
lib	2 minutes ago
kaggle_submission.csv	a minute ago
permutation_importance.csv	a minute ago
run_info.json	2 minutes ago
script_code_explainability.json	2 minutes ago
script_raw.ipynb	2 minutes ago
script.ipynb	2 minutes ago
script.out.json	2 minutes ago
script.png	2 minutes ago
script.py	2 minutes ago
test.pkl	a minute ago
training.pkl	2 minutes ago

Numeric to Scale

In the given input dataset there are **15 columns ** with numeric values as follows where we can convert those values to scale through log1p.

Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, 1stFlrSF, GrLivArea, TotRmsAbvGrd, GarageYrBlt, MoSold, YrSold, SalePrice

ML Explainability Tips:

The following codes are generated with a confidence level of **0.73** because of the following reasons apply to ** 15 ** column(s) that include ['Id', 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', '1stFlrSF', 'GrLivArea', 'TotRmsAbvGrd', 'GarageYrBlt', 'MoSold', 'YrSold', 'SalePrice'].

1. The imbalance score among target values of the columns, is **0.0** which is smaller or equal to **0.001**.
2. The maximum of skewness (according to scipy.stats.skew) among the columns is **24.452** which is greater than **1.078**.
3. The maximum of standard deviation of the normalized data (among the columns) is **0.344** which is smaller or equal to **0.38**.

```
[9]: # PREPROCESSING-4
# Component: Preprocess:Log
# Efficient Cause: Preprocess:Log is required in this pipeline since the dataset has ['feature:target_imbalance_score', 'feature:max_skewness', 'feature:max_normali
# Purpose: Return the natural logarithm of one plus the input array, element-wise.
# Form:
#   Input: array_like
#   Key hyperparameters used: None
# Alternatives: Although [Preprocess:StandardScaler] can also be used for this dataset, Preprocess:Log is used because it has more feature:target_imbalance_score th
# Order: Preprocess:Log should be applied...
import numpy as np
_NUMERIC_COLS_TO_SCALE = ['Id', 'MSSubClass', 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', '1stFlrSF', 'GrLivArea', 'TotRmsA
for _col in _NUMERIC_COLS_TO_SCALE:
    __train_dataset[_col] = np.log1p(__train_dataset[_col])
    if _col in __test_dataset:
        __test_dataset[_col] = np.log1p(__test_dataset[_col])
```

```
[10]: # DETACH TARGET
TARGET_COLUMNS = ['SalePrice']
if set(_TARGET_COLUMNS).issubset(__test_dataset.columns.tolist()):
```

You can see why each segment of the pipeline is generated and what its probability is by reviewing the code with its explanation.

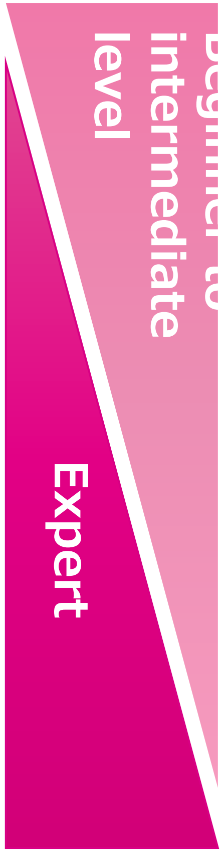
Train a Model

Information for proposal

- GH repo: https://github.com/sapientml/sapientml
- License: Apache 2.0
- Proposal: https://github.com/lfai/proposing-projects/blob/master/proposals/sapientml.adoc
- Possible Collaboration in LF AI&Data
  - Pre-processing: Amundsen, Feast, Feathr
  - Modelling: Adversarial Robustness Toolbox, AI Explainability 360, AI Fairness 360, Intersectional Fairness
  - Deployment: Acumos
  - Coding: Elyta, Kedro

The Value of SapientML

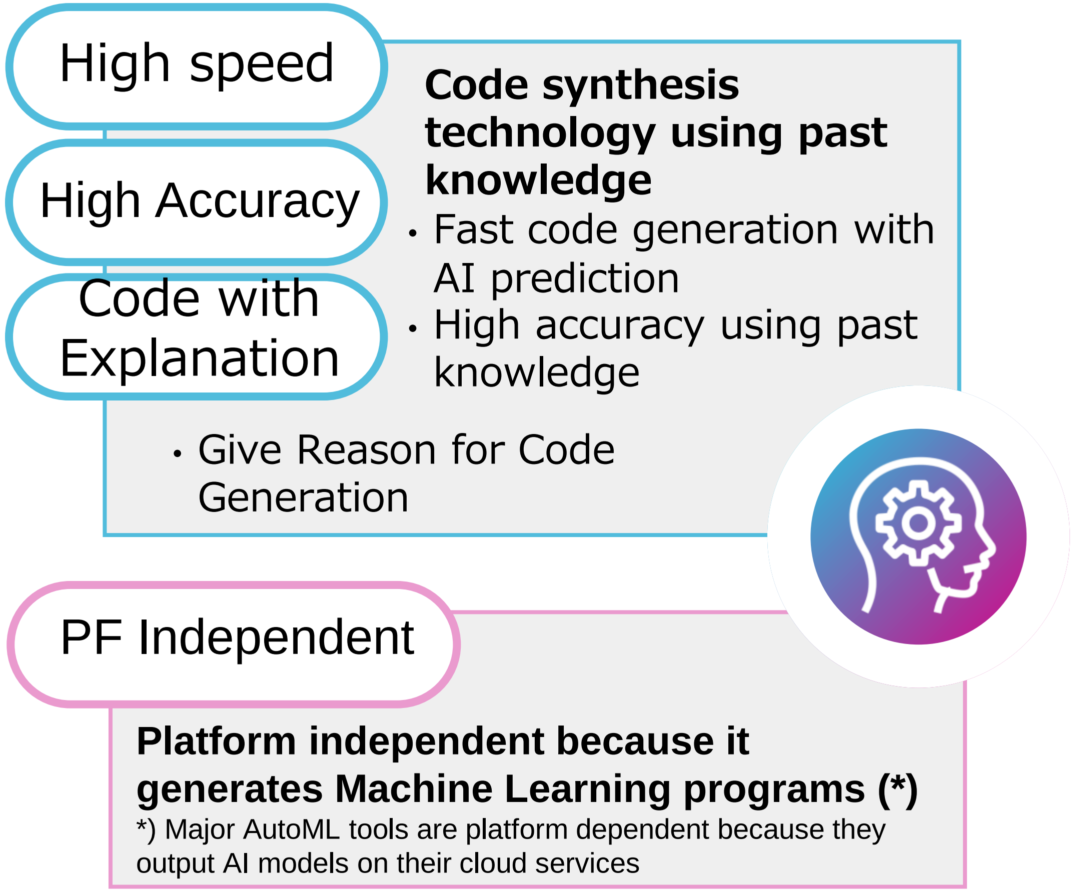
User
(Data Scientist)



Valu
e

- ✓ **No programming required. Generation of practical AI models**
- ✓ **Reduces the effort of trial and error by making clear model selection including code generation reasons**
- ✓ **Freely customizable including proprietary know-how and additional precision tuning**

SapientML Functions/Features



Code synthesis technology using past knowledge

- Fast code generation with AI prediction
- High accuracy using past knowledge
- Give Reason for Code Generation

Platform independent because it generates Machine Learning programs (*)

*) Major AutoML tools are platform dependent because they output AI models on their cloud services

Next Step

Next Step

- Support more meta-features of datasets
- Support more ML components
    - Preprocess components
    - Model

□ Accumulate more datasets and ML pipelines from community

Thank you

Approval on SapientML

Proposed Resolution:

- SapientML as a sandbox project of the LF AI & Data Foundation is hereby approved.

Upcoming TAC Meetings

LF AI & DATA

Upcoming TAC Meetings

- September 7 – Update from the Trusted AI committee, Update from the MLSecOps Committee
- September 21 – Marquez graduation request; AIDA, a new project requesting Sandbox Incubation

Please note we are always open to special topics as well.

If you have a topic idea or agenda item, please send agenda topic requests to tac-general@lists.lfaidata.foundation

Upcoming Events of Interest

- 2023 AICON Middle East Summit - October 8th to 9th in Riyadh https://lfaidata.foundation/blog/2023/07/18/2023-aicon-middle-east-summit-call-for-topics-from-around-the-world/
- Open Source Summit Europe in Bilbao, Spain, September 19-21 – LF AI&Data will have a booth https://events.linuxfoundation.org/open-source-summit-europe/

Open Discussion

LF AI & DATA

TAC Meeting Details

- To subscribe to the TAC Group Calendar, visit the wiki: https://wiki.lfaidata.foundation/x/cQB2 ______
- Join from PC, Mac, Linux, iOS or Android: https://zoom.us/j/430697670
- Or iPhone one-tap:
  - US: +16465588656,,430697670# or +16699006833,,430697670#
- Or Telephone:
  - Dial(for higher quality, dial a number based on your current location):
  - US: +1 646 558 8656 or +1 669 900 6833 or +1 855 880 1246 (Toll Free) or +1 877 369 0926 (Toll Free)
- Meeting ID: 430 697 670
- International numbers available: https://zoom.us/u/achYtcw7uN

Legal Notice

- The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at https://www.linuxfoundation.org/trademark-usage, as may be modified from time to time.
- Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at https://lmi.linuxfoundation.org for details regarding use of this trademark.
- Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.
- TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.
- Facebook and the "f" logo are trademarks of Facebook or its affiliates.
- LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.
- YouTube and the YouTube icon are trademarks of YouTube or its affiliates.
- All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.
- The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at https://www.linuxfoundation.org/privacy and its Antitrust Policy at https://www.linuxfoundation.org/antitrust-policy. each as may be modified from time to time. More information about The Linux Foundation's policies is available at https://www.linuxfoundation.org.
- Please email legal@linuxfoundation.org with any questions about The Linux Foundation's policies or the notices set forth on this slide.

LF AI & DATA

24AUG2023