goto;

GOTO
AARHUS 2021

#GOTOaar

Hello Qualification Set!

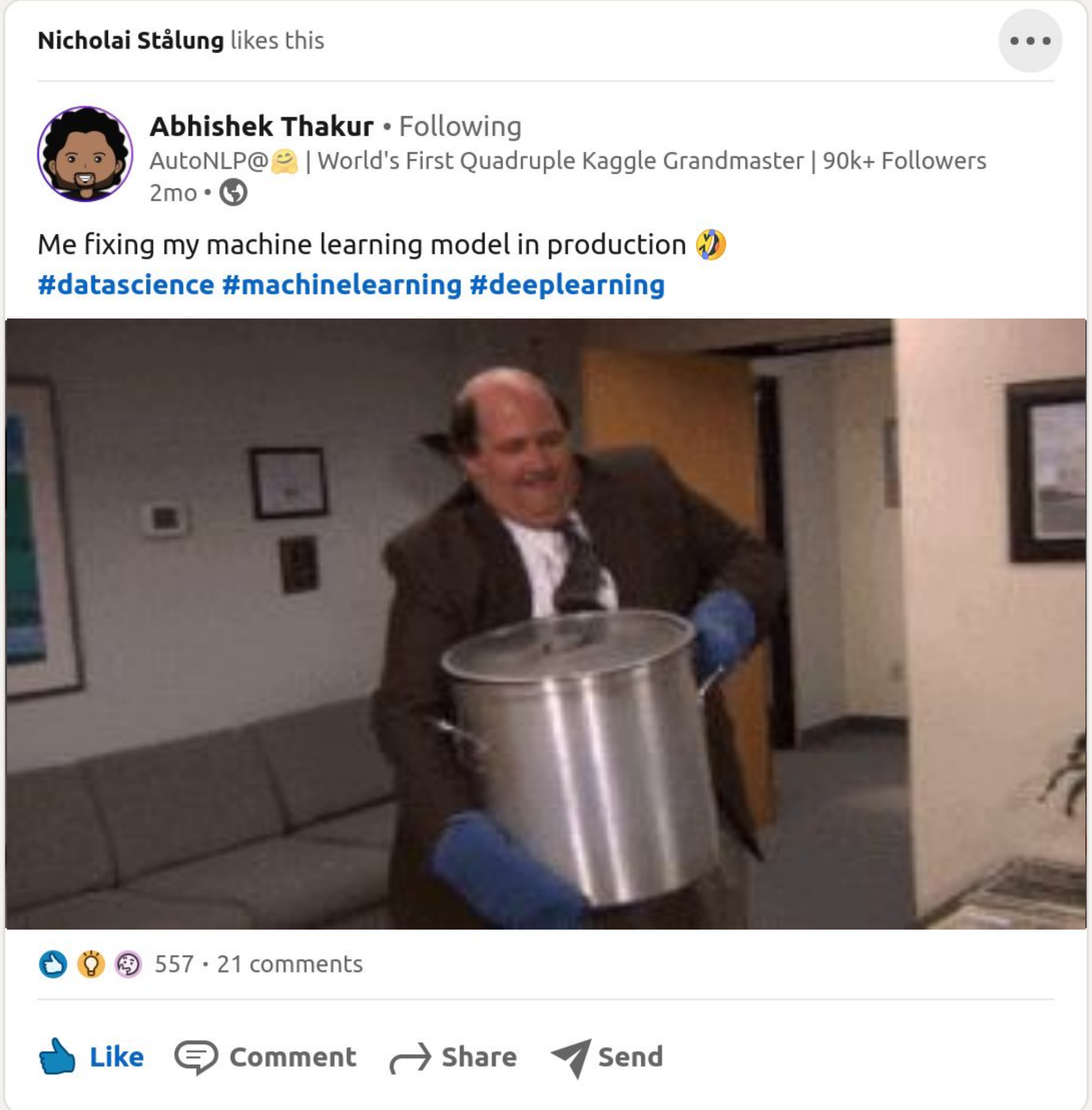
Nicholai Stålung, Lead Data Scientist, Trifork
+45 22598127
nrs@trifork.com

> “ All models are wrong, but some are useful.

George E. P. Box

Former president of the American Statistical Association

Problem?



Nicholai Stålung likes this

Abhishek Thakur • Following

AutoNLP@🤗 | World's First Quadruple Kaggle Grandmaster | 90k+ Followers

2mo •

Me fixing my machine learning model in production 🤣

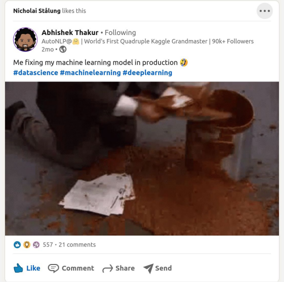
#datascience #machinelearning #deeplearning

557 • 21 comments

Overfitting

Data drift

Feedback loop



Nicholai Stålung likes this

Abhishek Thakur • Following
AutoNLP@🤗 | World's First Quadruple Kaggle Grandmaster | 90k+ Followers
2mo •

Me fixing my machine learning model in production 🤣
#datascience #machinelearning #deeplearning

557 • 21 comments

Accuracy < 90%

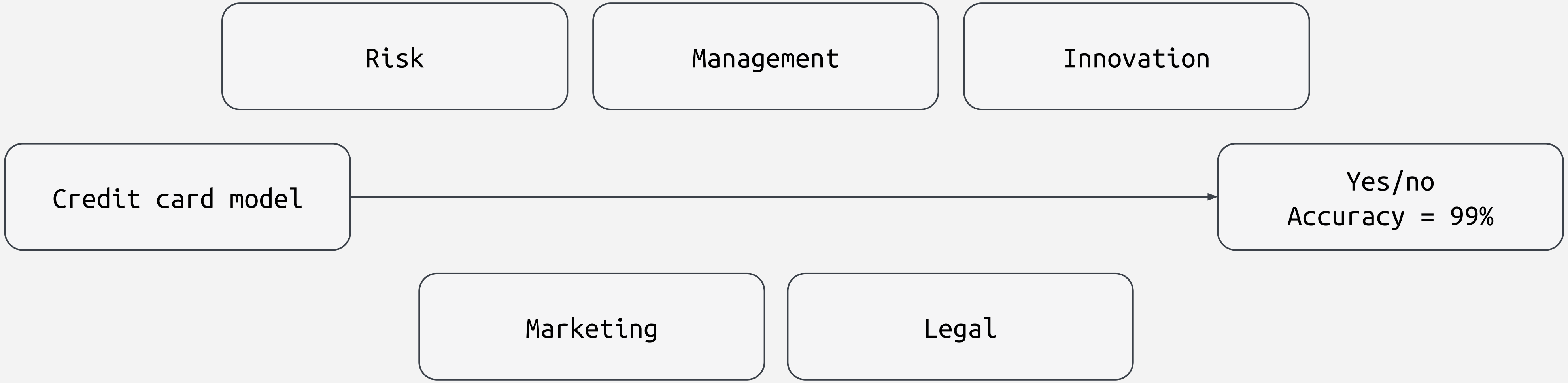
Undiscovered bias

Explainability

My first hypothesis?

Machine Learning won’t scale to critical applications on the cross-validation test

Financial institution



Risk

Management

Innovation

Credit card model

Yes/no
Accuracy = 99%

Marketing

Legal

My second hypothesis?

It’s not sustainable to test in production

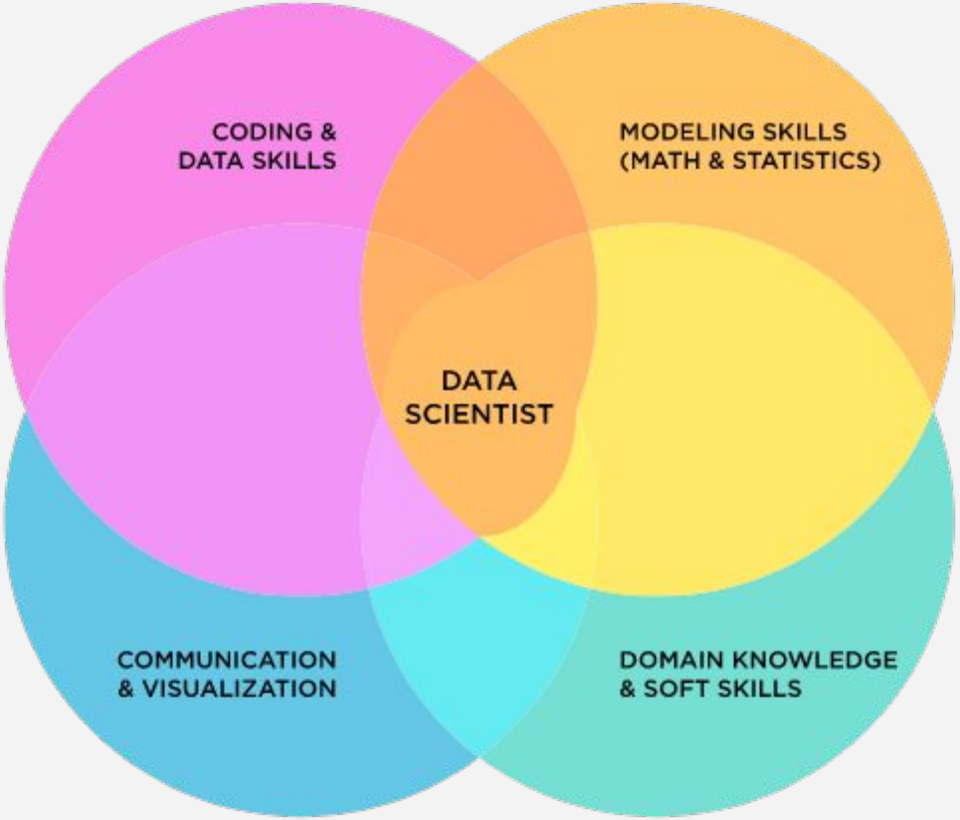
> “ 87 % of Data Science projects never make it into production.

https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/?ref=hackernoon.com

TRIFORK.

My third hypothesis?

Data Scientists shouldn’t verify their own systems


CODING & DATA SKILLS
MODELING SKILLS (MATH & STATISTICS)
DATA SCIENTIST
COMMUNICATION & VISUALIZATION
DOMAIN KNOWLEDGE & SOFT SKILLS

How do we continuously deploy models with ease, while keeping a high confidence in predictions?

Hello Qualification Set!

> “A qualification set is one or many controlled datasets used to qualify machine learning systems before deploying to production

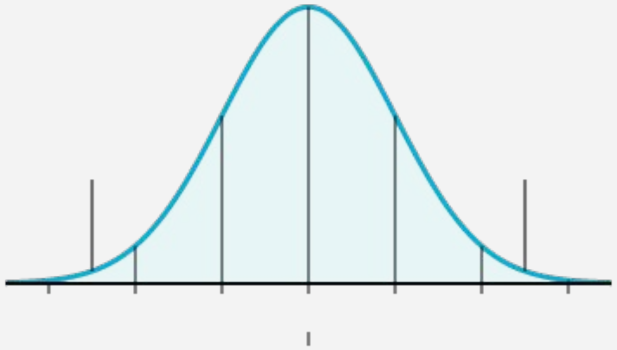
Qualify

- To have the necessary skill or knowledge to do a particular job or activity

https://www.merriam-webster.com/dictionary/qualify

Test set

- “Real” world data distributions
- A good indicator of the “normal” scenario
- We know that the tails are a problem



Qualification set

- Dynamic checklist
- Organizational task
- Corner-cases, testable observations and curiosity

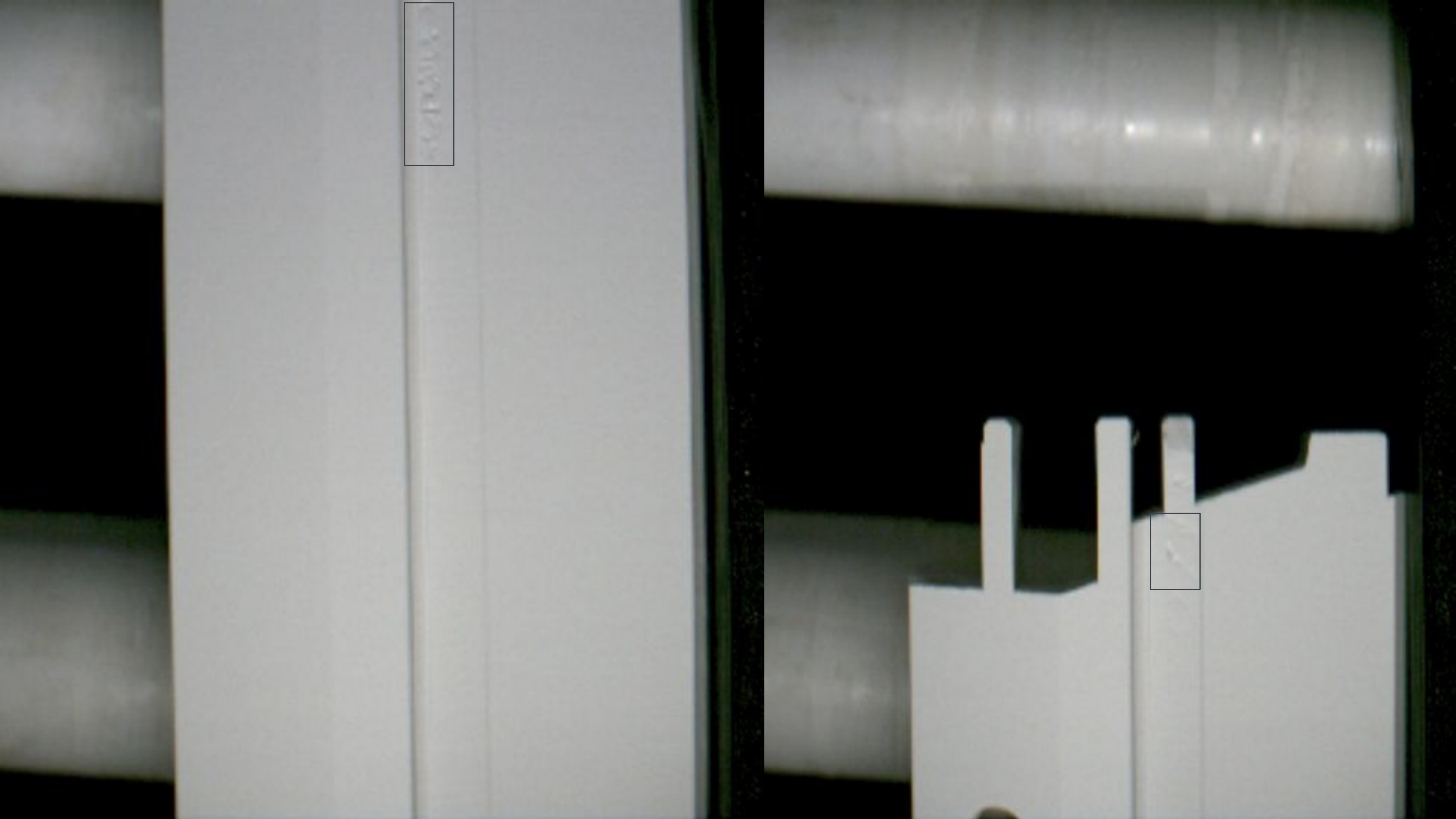
TRIFORK.

Creating a qualification set

- You need solution requirements!
  - Ask each stakeholder what they require from the system
  - Trivial cases
  - Nontrivial cases
- And you need a data map!
  - Define normal cases
  - Define abnormal cases
  - Define corner cases

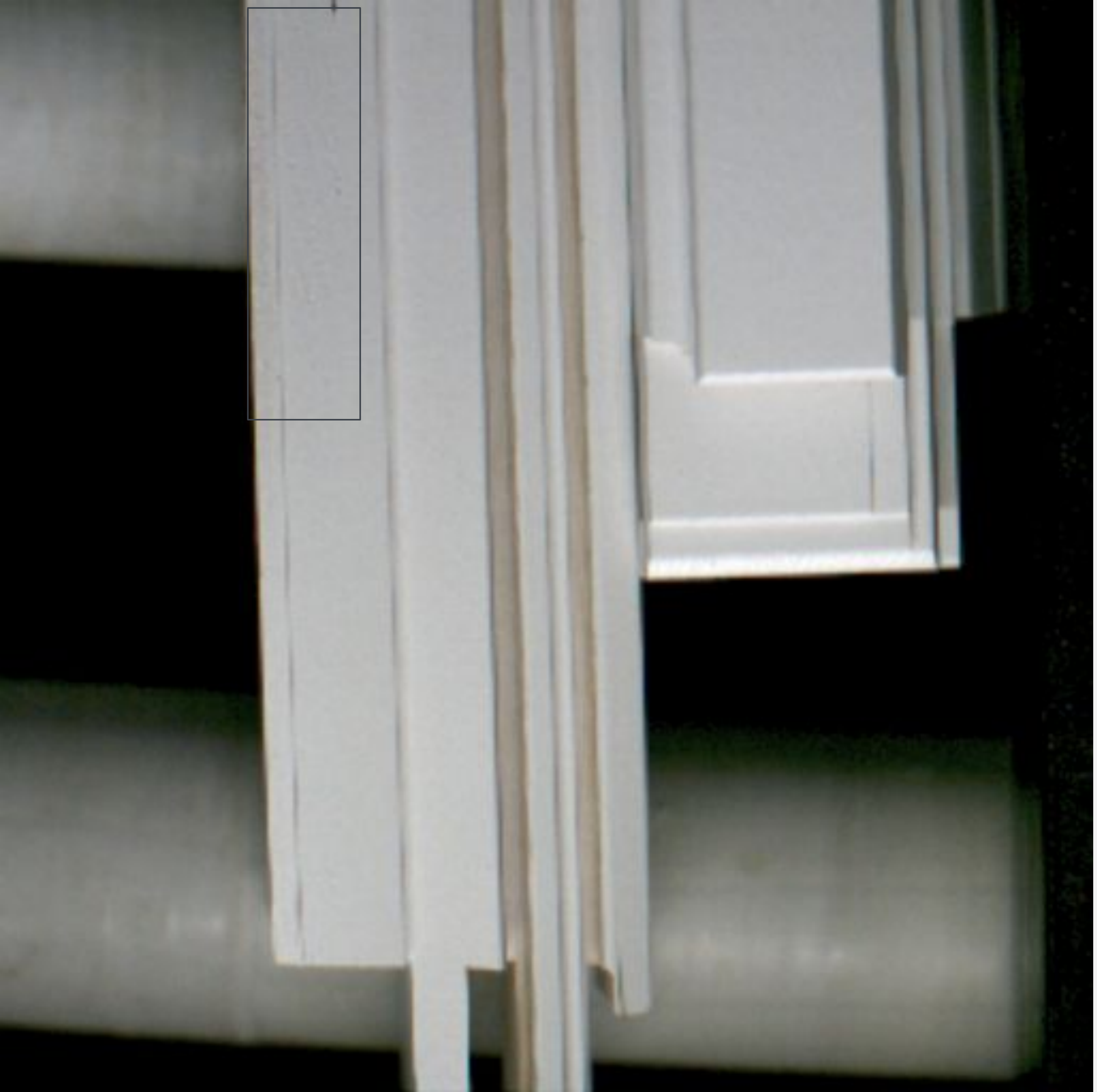
Real example of a Qualification set! - Manufacturing

- Solution requirements
  - Locate 95 % of all visual defects
  - Classify 80 % of defect type A correctly
- Data map
  - Identifier
  - Context description
  - Criticality
  - Expectations on requirements



-Type A
-Normal case
-Critical
-Expect to locate
-Expect to classify

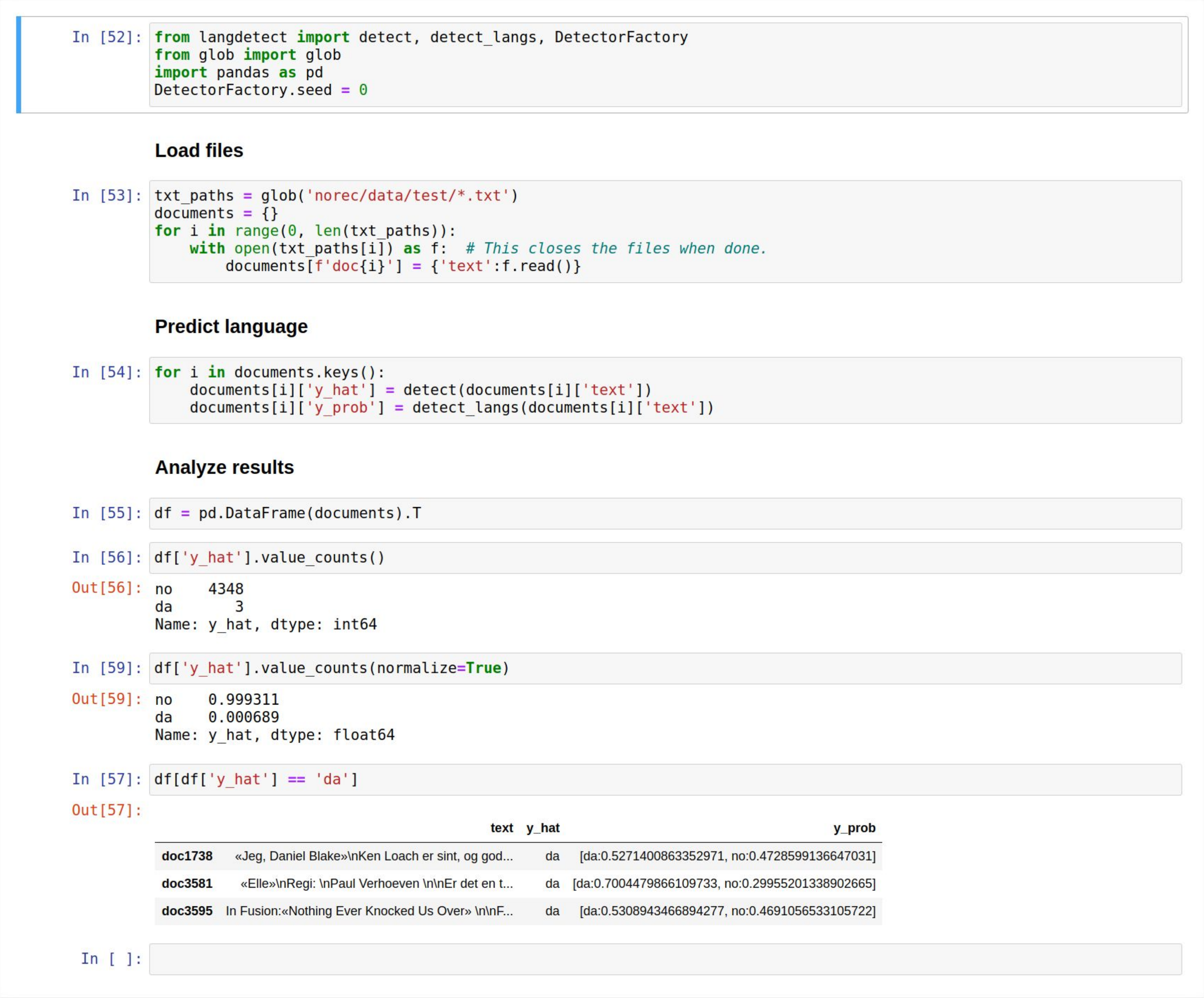
-Type A
-Abnormal case
-Non-critical
-Expect to locate
-Don’t expect to classify



-Type d
-Corner case
-Critical
-Don’t expect to locate
-Don’t expect to classify

Another example of a Qualification set! - Social listening

- Solution requirements
  - System should disregard norwegian texts
- Data map
  - Identifier



```
In [52]: from langdetect import detect, detect_langs, DetectorFactory
         from glob import glob
         import pandas as pd
         DetectorFactory.seed = 0
```

Load files

```
In [53]: txt_paths = glob('norec/data/test/*.txt')
         documents = {}
         for i in range(0, len(txt_paths)):
             with open(txt_paths[i]) as f:  # This closes the files when done.
                 documents[f'doc{i}'] = {'text':f.read()}
```

Predict language

```
In [54]: for i in documents.keys():
             documents[i]['y_hat'] = detect(documents[i]['text'])
             documents[i]['y_prob'] = detect_langs(documents[i]['text'])
```

Analyze results

```
In [55]: df = pd.DataFrame(documents).T
```

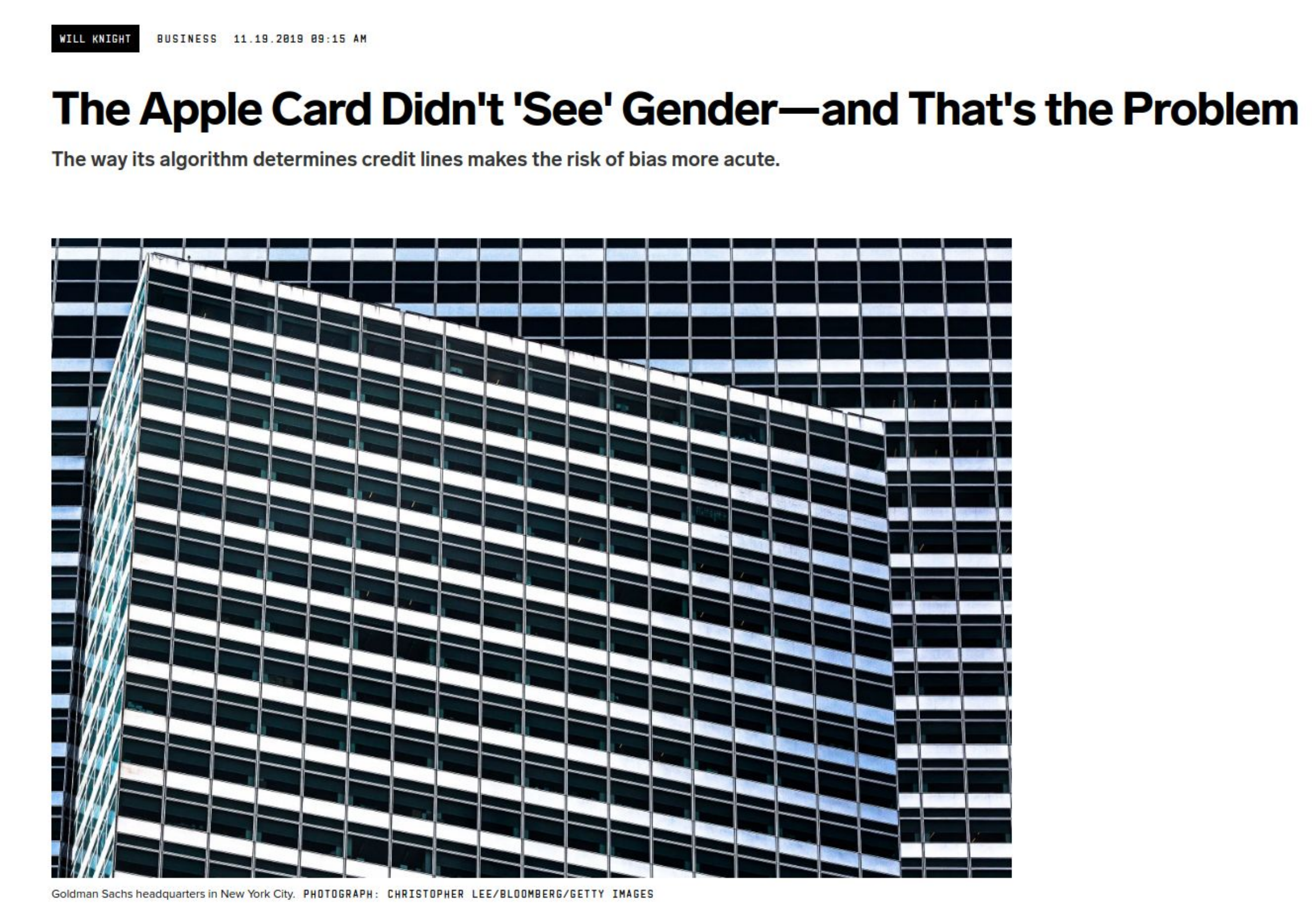
```
In [56]: df['y_hat'].value_counts()

Out[56]: no    4348
         da       3
         Name: y_hat, dtype: int64
```

```
In [59]: df['y_hat'].value_counts(normalize=True)

Out[59]: no    0.999311
         da    0.000689
         Name: y_hat, dtype: float64
```

```
In [57]: df[df['y_hat'] == 'da']

Out[57]:
```

	text	y_hat	y_prob
doc1738	«Jeg, Daniel Blake»\nKen Loach er sint, og god...	da	[da:0.5271400863352971, no:0.4728599136647031]
doc3581	«Elle»\nRegi: \nPaul Verhoeven \n\nEr det en t...	da	[da:0.7004479866109733, no:0.29955201338902665]
doc3595	In Fusion:«Nothing Ever Knocked Us Over» \n\nF...	da	[da:0.5308943466894277, no:0.4691056533105722]

```
In [ ]:
```



WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

The Apple Card Didn't 'See' Gender—and That's the Problem

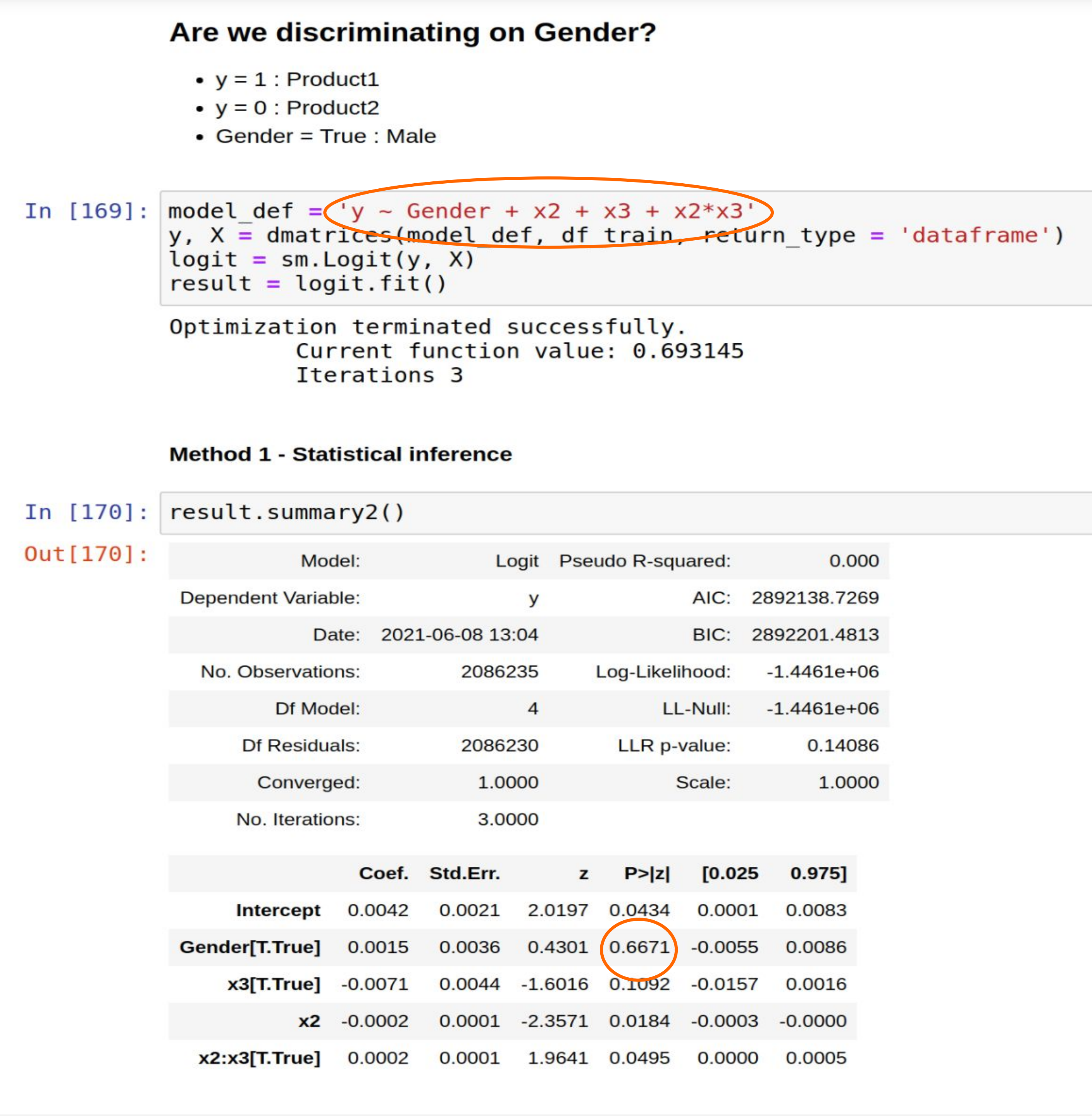
The way its algorithm determines credit lines makes the risk of bias more acute.

Goldman Sachs headquarters in New York City. PHOTOGRAPH: CHRISTOPHER LEE/BLOOMBERG/GETTY IMAGES

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Third example of a Qualification set! - Customer action model

- Solution requirements
    - System should not be biased towards gender
- Data map
    - ?



Are we discriminating on Gender?

- y = 1 : Product1
- y = 0 : Product2
- Gender = True : Male

In [169]:

```
model_def = 'y ~ Gender + x2 + x3 + x2*x3'
y, X = dmatrices(model_def, df_train, return_type = 'dataframe')
logit = sm.Logit(y, X)
result = logit.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.693145
         Iterations 3
```

Method 1 - Statistical inference

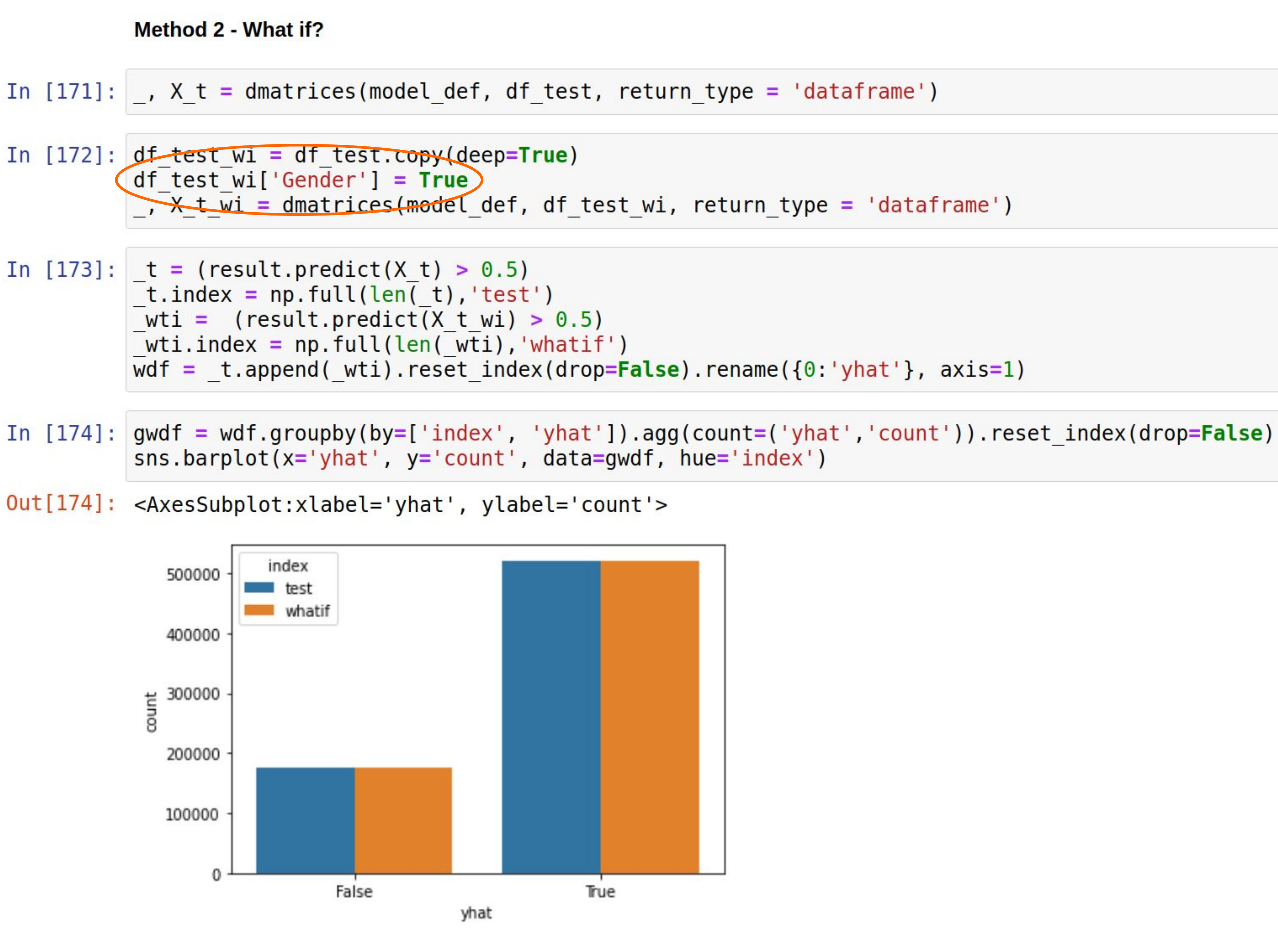
In [170]:

```
result.summary2()
```

Out[170]:

Model:	Logit	Pseudo R-squared:	0.000
Dependent Variable:	y	AIC:	2892138.7269
Date:	2021-06-08 13:04	BIC:	2892201.4813
No. Observations:	2086235	Log-Likelihood:	-1.4461e+06
Df Model:	4	LL-Null:	-1.4461e+06
Df Residuals:	2086230	LLR p-value:	0.14086
Converged:	1.0000	Scale:	1.0000
No. Iterations:	3.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.0042	0.0021	2.0197	0.0434	0.0001	0.0083
Gender[T.True]	0.0015	0.0036	0.4301	0.6671	-0.0055	0.0086
x3[T.True]	-0.0071	0.0044	-1.6016	0.1092	-0.0157	0.0016
x2	-0.0002	0.0001	-2.3571	0.0184	-0.0003	-0.0000
x2:x3[T.True]	0.0002	0.0001	1.9641	0.0495	0.0000	0.0005



Method 2 - What if?

```
In [171]: _, X_t = dmatrices(model_def, df_test, return_type = 'dataframe')
```

```
In [172]: df_test_wi = df_test.copy(deep=True)
          df_test_wi['Gender'] = True
          _, X_t_wi = dmatrices(model_def, df_test_wi, return_type = 'dataframe')
```

```
In [173]: _t = (result.predict(X_t) > 0.5)
          _t.index = np.full(len(_t),'test')
          _wti =  (result.predict(X_t_wi) > 0.5)
          _wti.index = np.full(len(_wti),'whatif')
          wdf = _t.append(_wti).reset_index(drop=False).rename({0:'yhat'}, axis=1)
```

```
In [174]: gwdf = wdf.groupby(by=['index', 'yhat']).agg(count=('yhat','count')).reset_index(drop=False)
          sns.barplot(x='yhat', y='count', data=gwdf, hue='index')
```

```
Out[174]: <AxesSubplot:xlabel='yhat', ylabel='count'>
```

Key takeaways - Qualification sets

- Tool to certify systems in line with ither software certifications
- Can be used to verify model performance for critical/non-trivial situations
- Each set should answer one question
- Usable in supervised, unsupervised and pretrained models
- Can be used for all data formats
- Allows stakeholders to take ownership

goto;

Don’t forget to
vote for this session
in the **GOTO Guide app**