EXPECTATIONS VS. REALITY

MATTHIAS AND NIKOS WHO ARE THEY?

AND WHAT WILL THEY DO TODAY?

NIKOS LEKKAS

SENIOR DATA SCIENTIST

At Deutsche Telekom since May 2020.
Manager of Analytics Team at Dixons SE Europe 2016-2020
Senior Consultant at IRI 2010-2016
Analyst at VPRC doing political and social research 2007-2009

https://www.linkedin.com/in/nikos-lekkas-10894966/

MATTHIAS KRUMREIN

SENIOR CONSULTANT

ODISYS
we love. we do.

- As a consultant active for ~ 15 years
- With focus on Business Intelligence (BI)
  - Data Analyst
  - Data Engineer
  - Data Scientist

Bachelor Thesis | Master Thesis | Jobs

https://www.odisys.de/jobs

https://www.linkedin.com/in/matthias-krumrein-791a251ab/

OUTLINE

- What is ML quick intro
- Expectations vs Reality from a business perspective
- A practical example of predicting the probability of selling a product

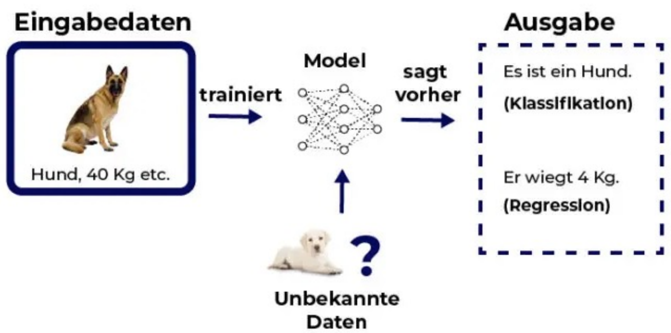
WHAT IS MACHINE LEARNING (ML)

Machine learning usually refers to the **changes in systems** that perform tasks associated with **artificial intelligence (AI)**.

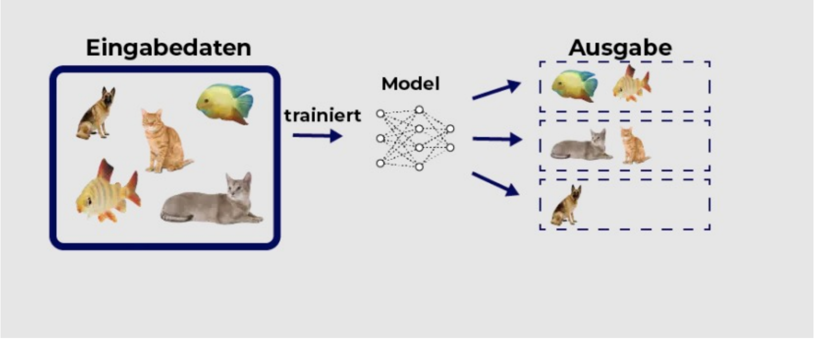
An artificial system **learns from examples** and can **generalise them** once the learning phase is complete. To do this, machine learning **algorithms build a statistical model** based on **training data**.

> **When the performance of a speech-recognition machine improves** after hearing several samples of a person's speech, […] in that case to say **that the machine has learned**.

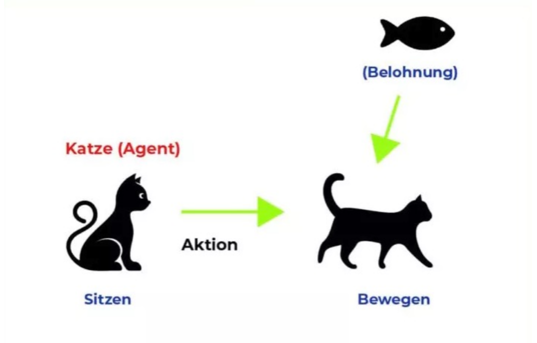
Examples are not simply learned by heart, but patterns and regularities are recognised in the learning data
A ML system can also assess unknown data



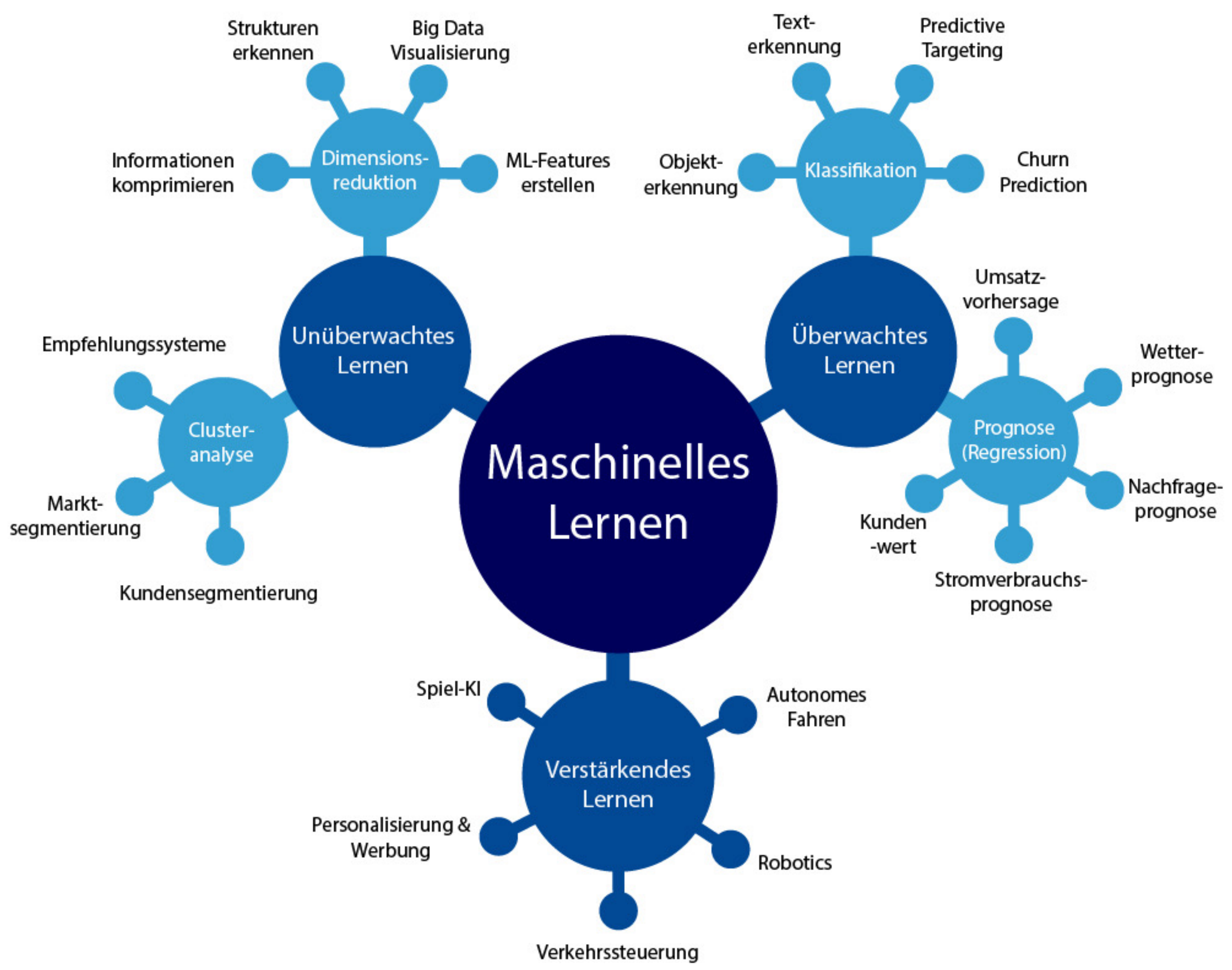
- Supervised Learning



- Unsupervised Learning

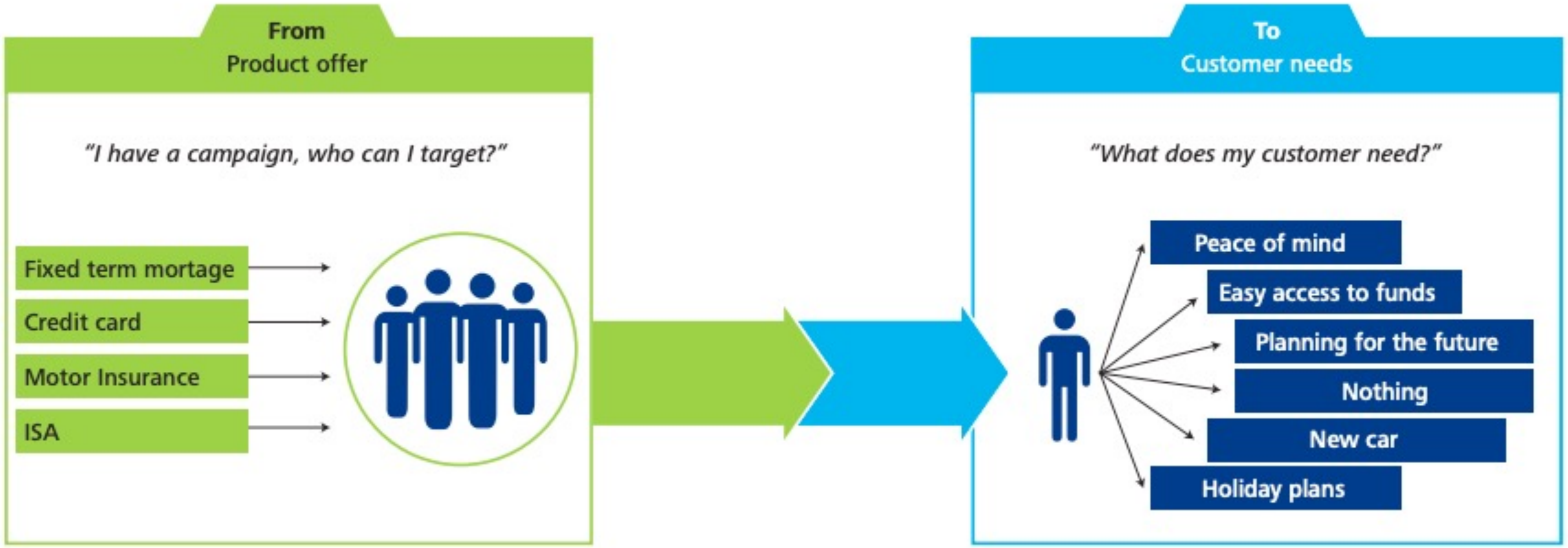


- Reinforcement Learning


Maschinelles Lernen
Unüberwachtes Lernen
Dimensions-reduktion
Strukturen erkennen
Big Data Visualisierung
Informationen komprimieren
ML-Features erstellen
Cluster-analyse
Empfehlungssysteme
Markt-segmentierung
Kundensegmentierung
Überwachtes Lernen
Klassifikation
Text-erkennung
Predictive Targeting
Objekt-erkennung
Churn Prediction
Prognose (Regression)
Umsatz-vorhersage
Wetter-prognose
Nachfrage-prognose
Kunden -wert
Stromverbrauchs-prognose
Verstärkendes Lernen
Spiel-KI
Autonomes Fahren
Personalisierung & Werbung
Robotics
Verkehrssteuerung

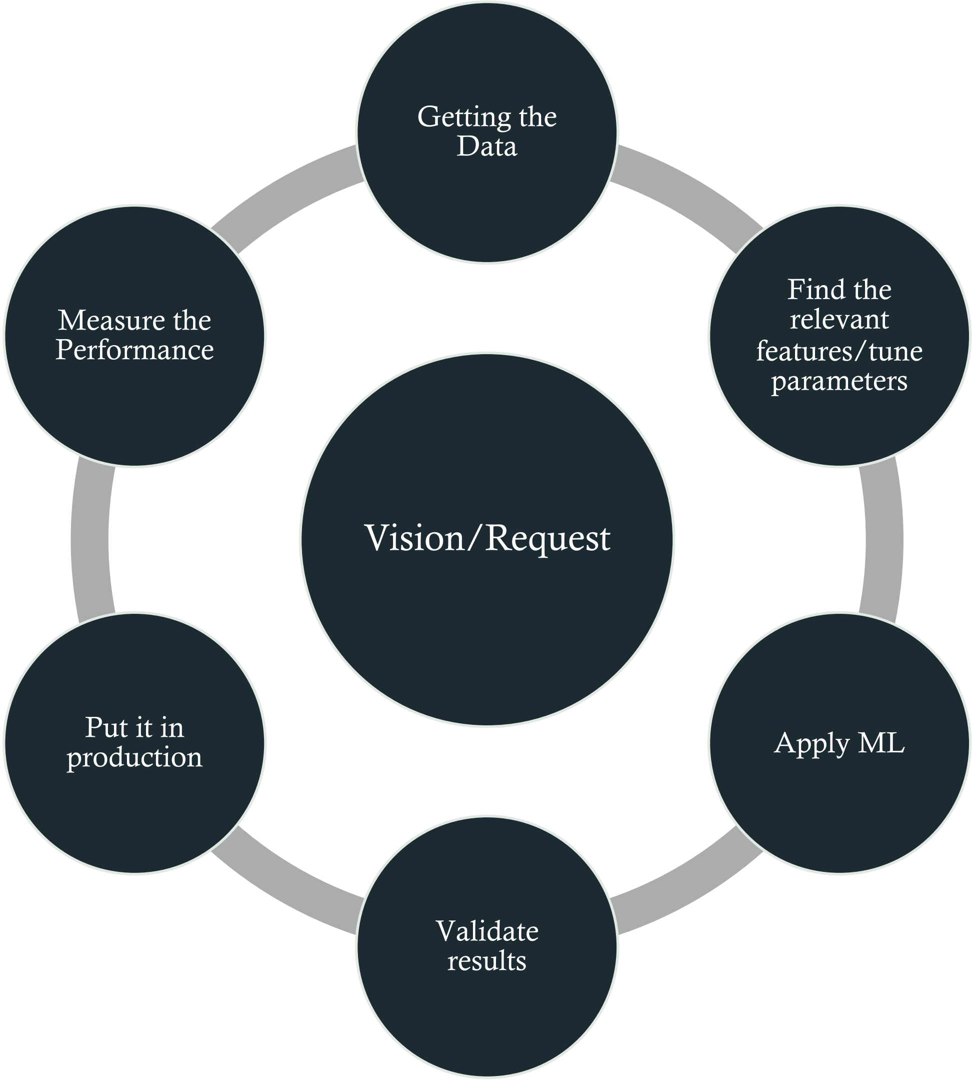
Next best action (NBA)

NBA recommendations can be improved through more **accurate data, analytics and value calculations**. **Predicting** each customer's **likelihood** to respond **to certain products** allows a cost-effective approach to marketing, as **resources are not wasted** on offers that are likely to be rejected



https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consultancy/deloitte-uk-con-next-best-action.pdf

WHAT ARE THE NEEDED STEPS IN A PERFECT WORLD TO IMPLEMENT ML SOLUTION IN A BUSINESS PROCESS


Getting the Data
Find the relevant features/tune parameters
Apply ML
Validate results
Put it in production
Measure the Performance
Vision/Request

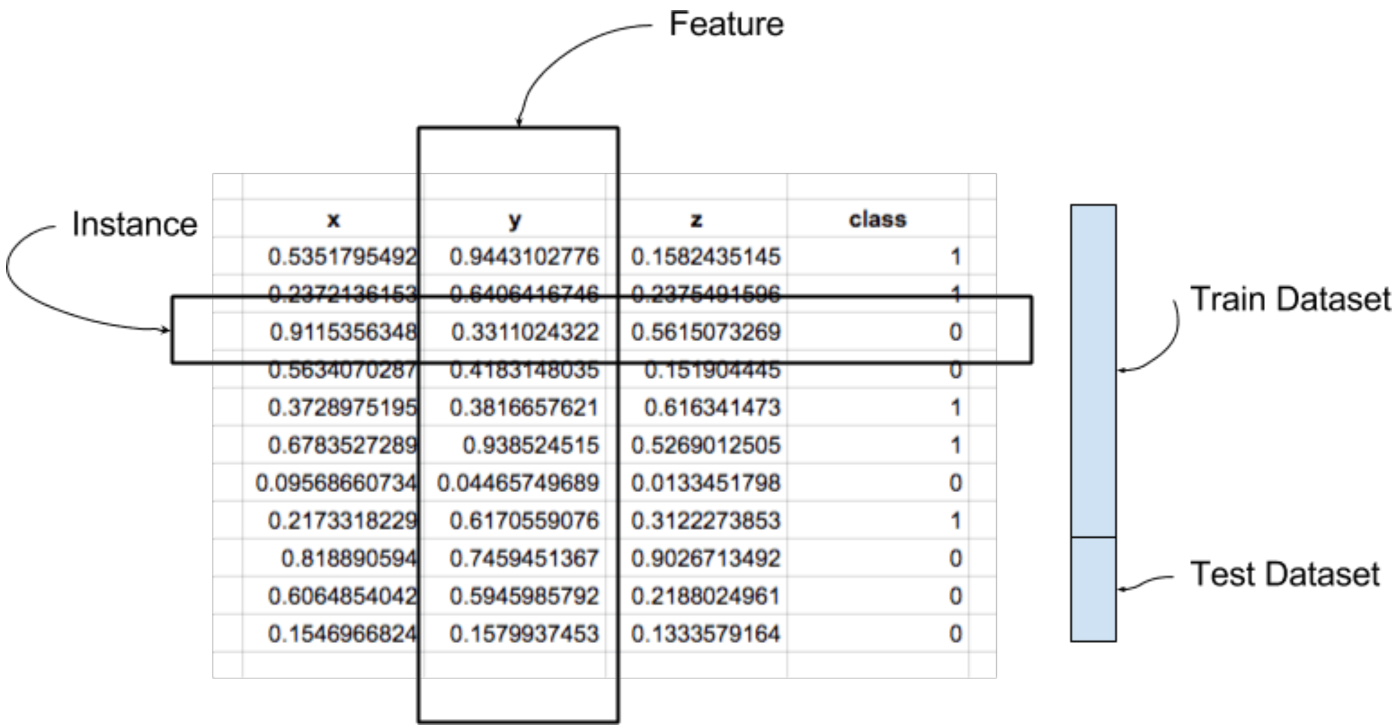
DO STAKEHOLDERS HAVE A PLAN?

WHAT IS THE PLAN

EXPECTATION	REALITY
“I have some data! Just push the play button and let the ML Algorithms do their work. Next month we will have doubled our sales”	“These data are not suitable for running an ML algorithm. We need to narrow down the problem”
“Bring us a list of our best 10 customers which are willed to spend a lot more money”	“Yes but the only thing we know about the customers is their Postal Code”
“You must predict the age of our customers”	“Okay, but we have only the customer ID”
„Get to pick the problem to work on“	“Problems are driven by business need“

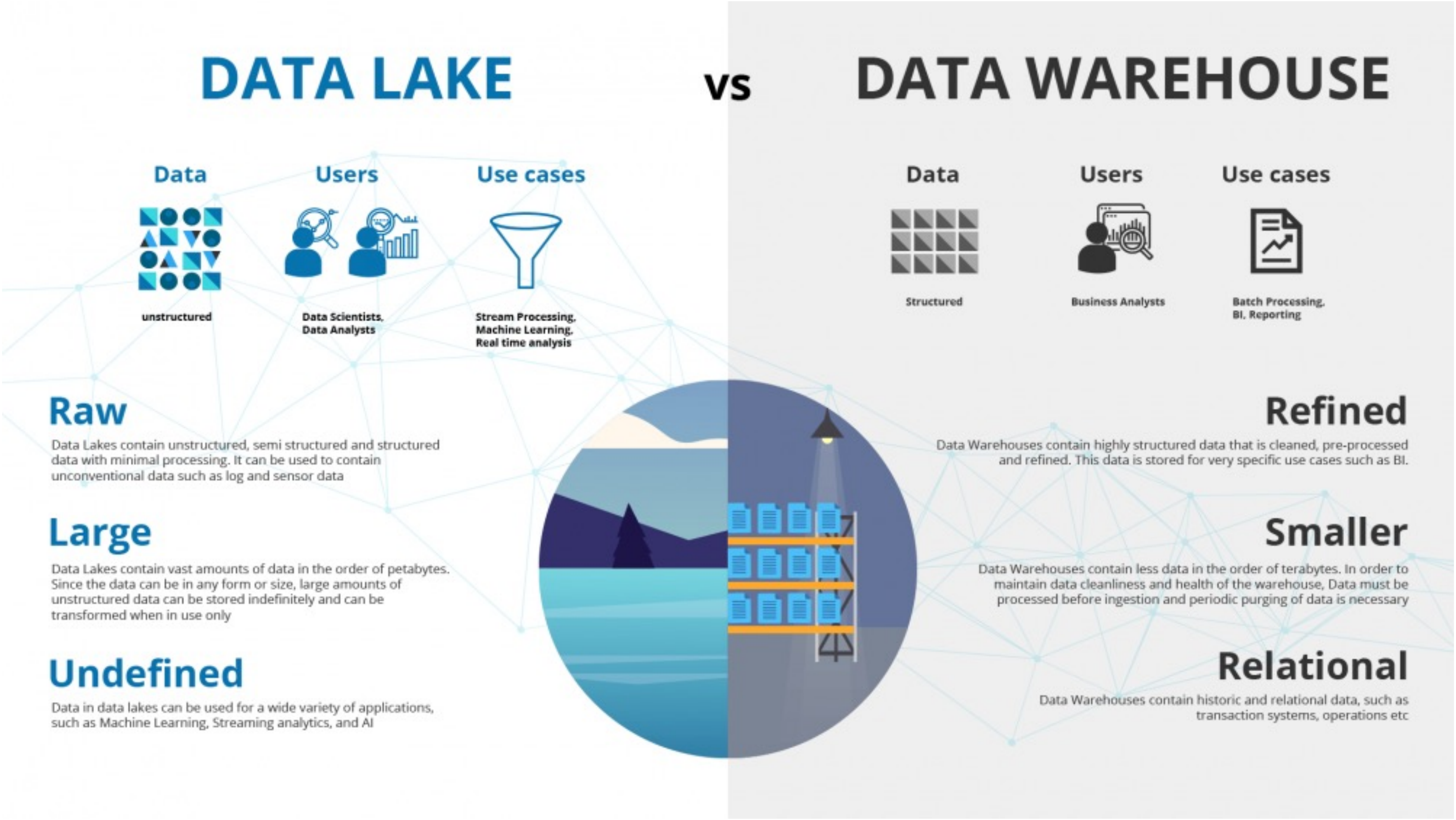
DO WE HAVE THE DATA?

DO WE HAVE THE “RIGHT” DATA?



x	y	z	class
0.5351795492	0.9443102776	0.1582435145	1
0.2372136153	0.6406416746	0.2375491596	1
0.9115356348	0.3311024322	0.5615073269	0
0.5634070287	0.4183148035	0.151904445	0
0.3728975195	0.3816657621	0.616341473	1
0.6783527289	0.938524515	0.5269012505	1
0.09568660734	0.04465749689	0.0133451798	0
0.2173318229	0.6170559076	0.3122273853	1
0.818890594	0.7459451367	0.9026713492	0
0.6064854042	0.5945985792	0.2188024961	0
0.1546966824	0.1579937453	0.1333579164	0

- Data are often distributed and have to be processed and merged.
- The quality of the data is often insufficient.



„Start supervised learning with a wealth of historic data […] Start with clean data, not big data" . - *Danny Lange*

EXPECTATION | REALITY

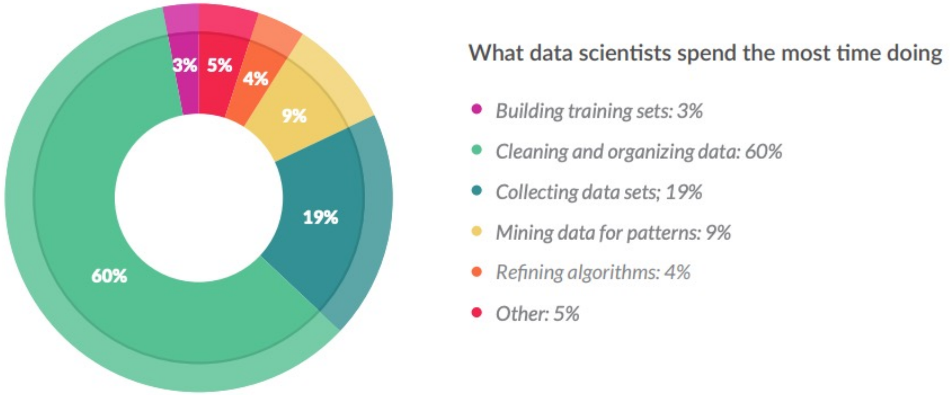
"Standard benchmark datasets, often with fixed features"

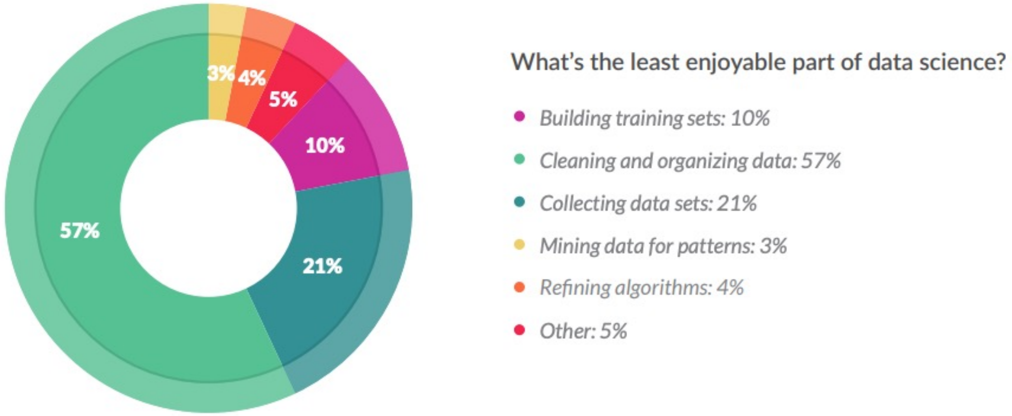
"We have all the relevant features for the last 5 years and we can easily access them"

"Since all available data come from our business system and have a structured format can be used immediately for our model"

"All needed tooling and infrastructure for processing and accessing the data is available"

„Novel, never-before-seen datasets with lots of space for feature engineering"





Data Science Report 2016 - https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

NOW RUN THE AUTO ML MAGIC

EXPECTATION	REALITY
"Since I'm using this sophisticated ML software the output should be correct"	"You need domain knowledge and a general understanding of different ML approaches to get started."
"The Software is support ML out of the Box – so everything is there just put in the data"	"Software is the tooling and will not show you automatically the right way forward"
"Which are the most important reasons for causing loss of sales? "	"Sometimes you need to sacrifice accuracy and complicated models if you cannot answer simple business questions"
"ML can solve all of my problems since is a sophisticated and state of the art process"	ML often fails. It's not the general answer of everything. It all depends on how well you'll define your problem.

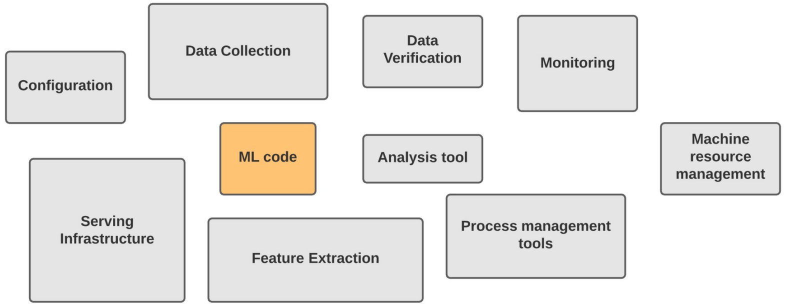
IS IT FINISHED AND WORKING ALREADY?

EXPECTATION

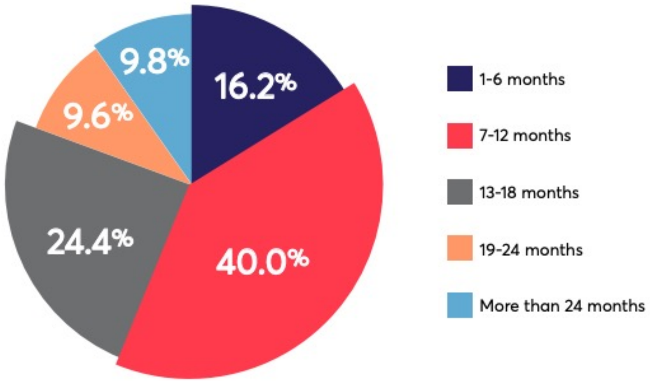
REALITY

“ Running a good model is just a small portion of the whole process”

“So the model is finished in round about 5 hours, so we are ready to go?”



“All Data is there you have good laptop so we expect that we can use the ML model for increasing our sales next month”



https://dotscience.com/assets/downloads/Dotscience_Survey-Report-2019.pdf

WE RUN IT ONCE – IT SEEMS TO WORK – NOW WHAT?

“SALES IS COMPLAINING IT IS ALL WRONG!”

EXPECTATION	REALITY

"It worked perfect on test data. So, it's easy to use it in production – it will work perfect with real data"

"Everyone is happy to get ML supported decisions and will use it"

"Easy to trust the numbers, they telling you always the truth"

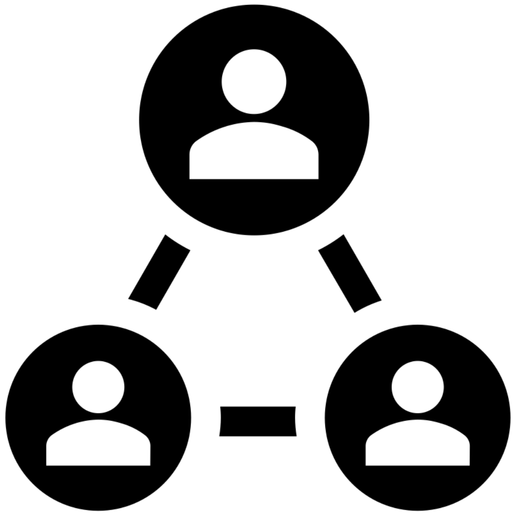
"I will always promote data analysis because I'm into the numbers, but the human factor will always matter. When we talk about fashion trends, an algorithm can't give a gut reaction, an algorithm can't go to Paris, London and Milan and say 'Oh my God, I know this is perfect for my customer'. Because algorithms only rely on historical data and oftentimes you need to be able to use that sixth sense that you have as a merchant, that tells you
[…] That's why you need smart buyers and visionary merchants in retailing"

A.S. Senior Buyer SDA

ARE THESE TWO GUYS JUST DOING FUN OF IT?

LET’S LOOK AT A REAL PROJECT

PREDICTING THE PROBABILITY OF PURCHASING A SPECIFIC PRODUCT (SIMPLE NBA)



- Supporting sales by preselecting customer with a high probability of purchasing a specific product

- Team skills: “Data Engineers, Data Scientists and Sales”
- Team knowledge: “Deep knowledge of available data and general business context”

STEP 1: FOR RUNNING A ML PROJECT
SELECT THE LEVEL OF AGGREGATION

- Decide for the level of analysis. (Customer ,customer-product level, customer-product-store). Depending on your level you'll have to do the appropriate aggregation

	created_on	company_id	company_entitlement_id	status	application_id	name	order_price	rank_count	rank_count_teams	is_free_app	...	tot
0	2012-07-03 14:58:29	2	863.0	CANCELLED	3.0	QA_TEST_1	NaN	1.0	0.0	1.0	...	
1	2012-07-05 17:01:53	2	916.0	CANCELLED	195.0	STRATO HiDrive Business	NaN	2.0	0.0	0.0	...	
2	2012-07-10 08:54:58	2	1026.0	CANCELLED	226.0	HiDrive Pro 2500	NaN	3.0	0.0	0.0	...	
3	2012-07-11 14:38:05	2	1125.0	CANCELLED	193.0	QA_TEST_2	NaN	4.0	0.0	0.0	...	
4	2012-07-16 14:24:58	2	1524.0	CANCELLED	227.0	HiDrive Pro 1000	NaN	5.0	0.0	0.0	...	

5 rows × 30 columns

STEP 1: FOR RUNNING A ML PROJECT

SELECT THE LEVEL OF AGGREGATION

- Decide for the level of analysis. (Customer ,customer-product level, customer-product-store). Depending on your level you'll have to do the appropriate aggregation

	created_on	company_id	company_entitlement_id	status	application_id	name	order_price	rank_count	rank_count_teams	is_free_app	...	tot
0	2012-07-03 14:58:29	2	863.0	CANCELLED	3.0	QA_TEST_1	NaN	1.0	0.0	1.0	...	
1	2012-07-05 17:01:53	2	916.0	CANCELLED	195.0	STRATO HiDrive Business	NaN	2.0	0.0	0.0	...	
2	2012-07-10 08:54:58	2	1026.0	CANCELLED	226.0	HiDrive Pro 2500	NaN	3.0	0.0	0.0	...	
3	2012-07-11 14:38:05	2	1125.0	CANCELLED	193.0	QA_TEST_2	NaN	4.0	0.0	0.0	...	
4	2012-07-16 14:24:58	2	1524.0	CANCELLED	227.0	HiDrive Pro 1000	NaN	5.0	0.0	0.0	...	

5 rows × 30 columns

STEP 2: FOR RUNNING A ML PROJECT

SELECT THE RIGHT FEATURES

Select the appropriate features after doing some research or discussing the problem with some key stakeholders. Sometimes you need to be creative and build new features from the existing ones

List of features: *Customer Ageing*, Previous purchases per application, Number of subscription licenses, *number of total applications*, Voucher Dummy

The magic of domain expertise: Reach out to domain experts, talk to them about features you can often get 80% of the benefit of deep learning models with about 10% of the effort

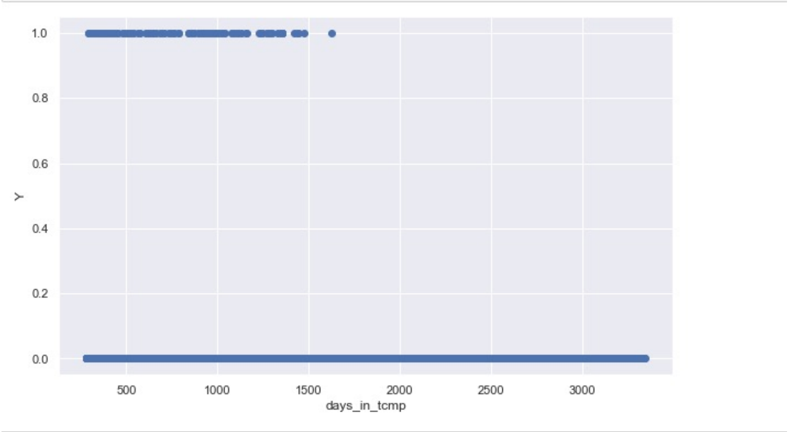
STEP 3: FOR RUNNING A ML PROJECT
RUN SOME EXPLORATORY DATA ANALYSIS (EDA)

- In real business problem you'll often face the issue of significant imbalances. The latter means that the amount of data that the model will use to predict is too few to make accurate predictions

```
Out[45]: 0    233651
         1       137
         Name: Y, dtype: int64
```

Non- Buyers

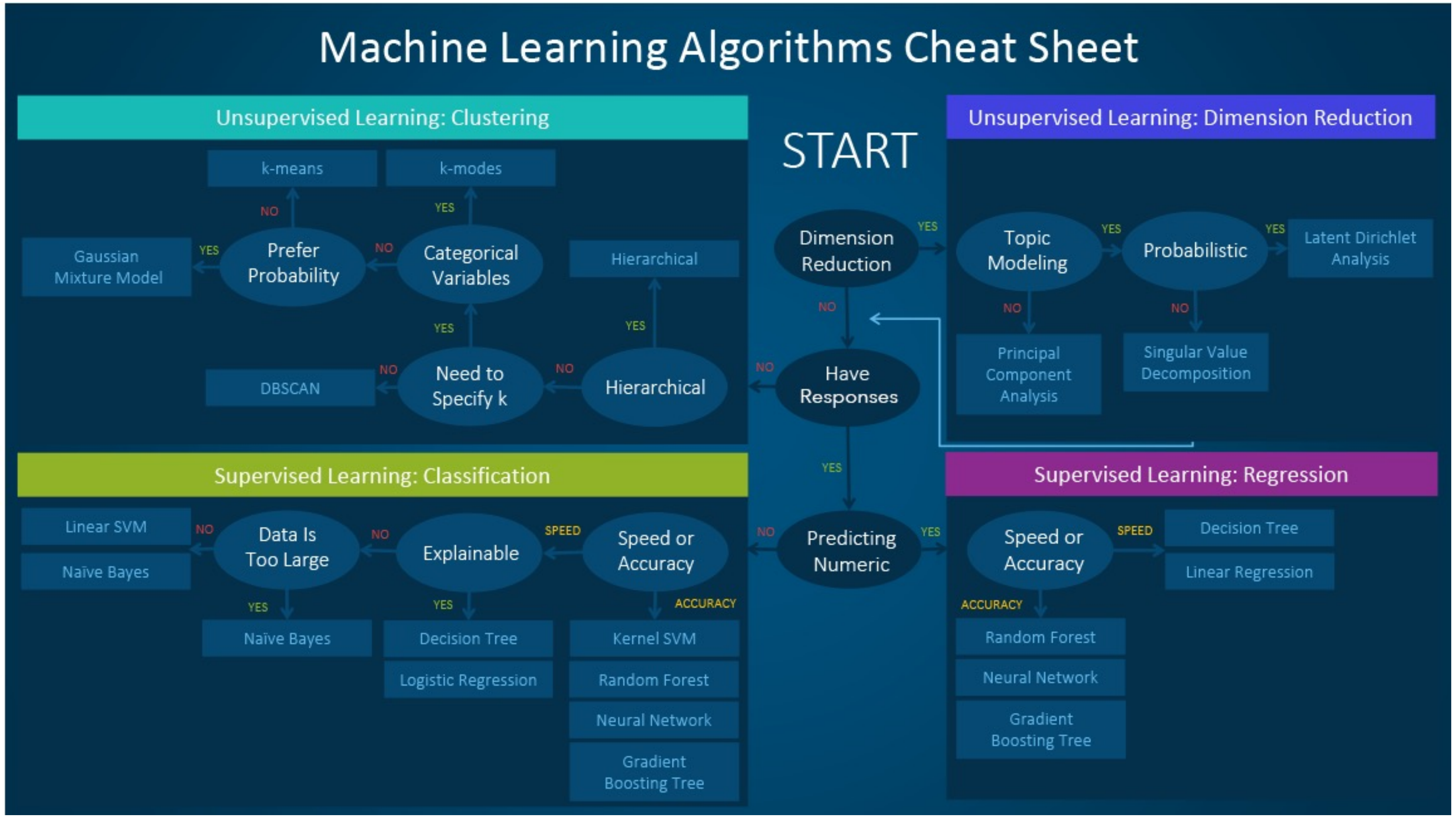
Buyers



STEP 4: FOR RUNNING A ML PROJECT

RUN DIFFERENT ML ALGORITHMS AND SELECT THE ONE THAT PERFORMS BETTER.

- Depending on the problem you would like to solve or the kind of data that you are using each algorithm will perform differently.

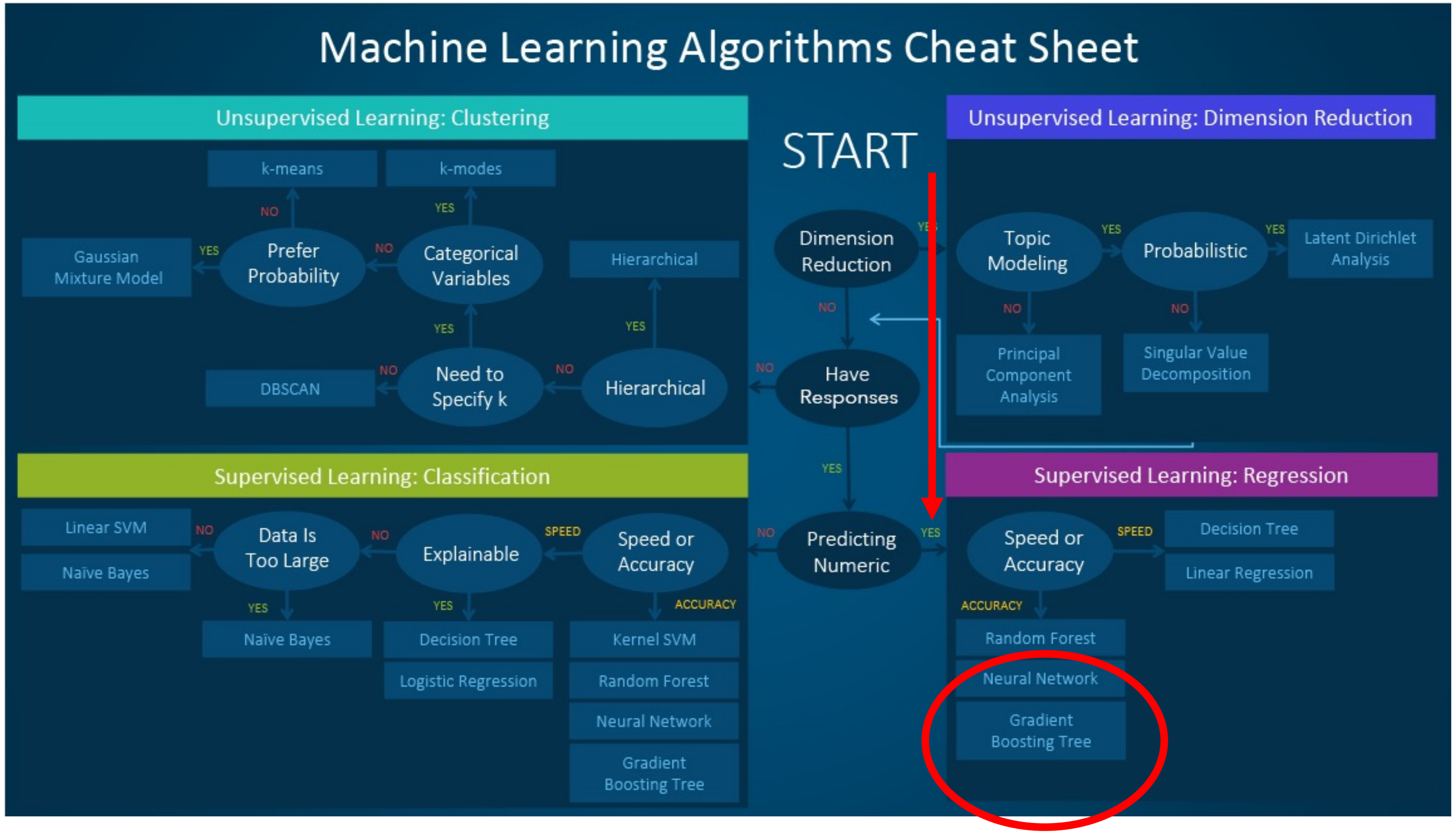


https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

STEP 4: FOR RUNNING A ML PROJECT

RUN DIFFERENT ML ALGORITHMS AND SELECT THE ONE THAT PERFORMS BETTER.

- Depending on the problem you would like to solve or the kind of data that you are using each algorithm will perform differently.

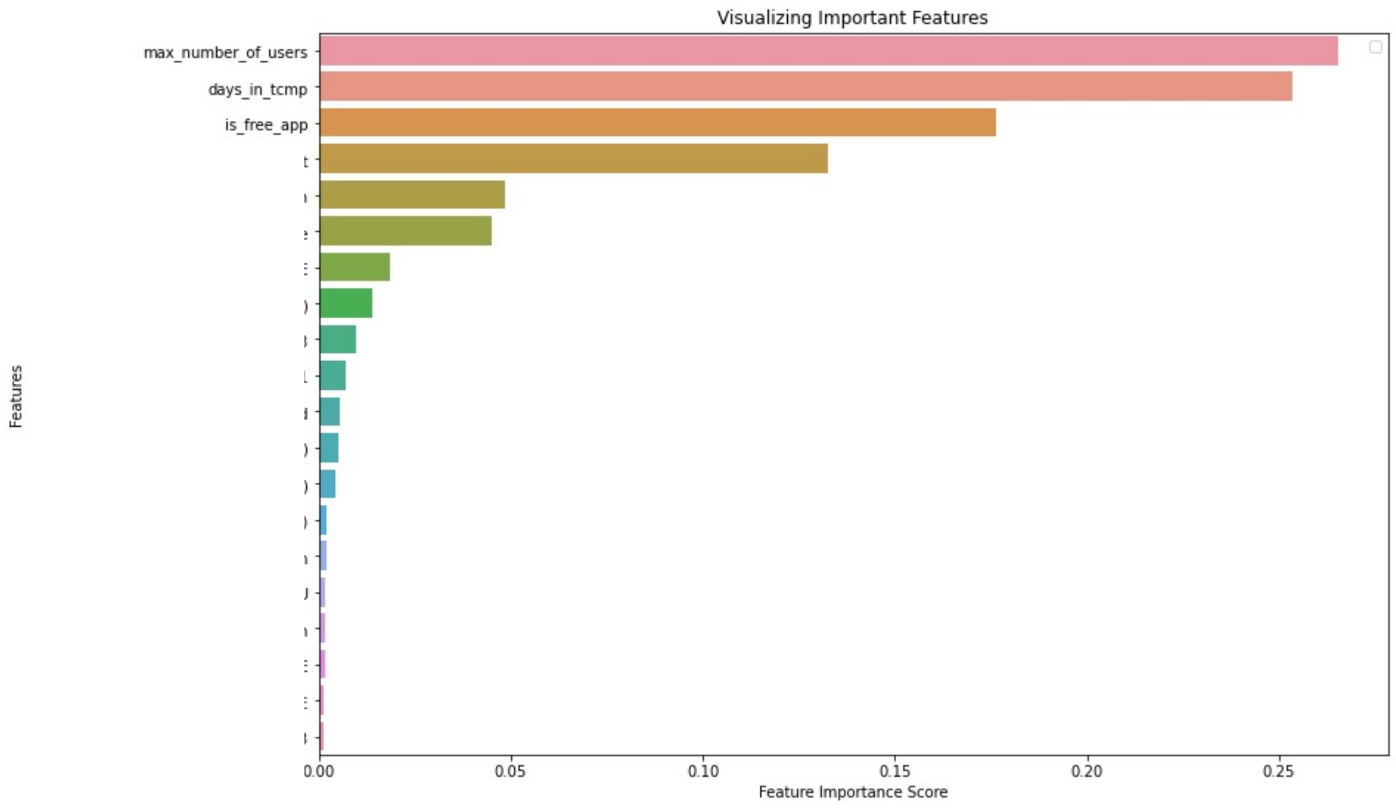


https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

STEP 5: FOR RUNNING A ML PROJECT

CHECK IF THE SIGNIFICANT PARAMETERS MAKE SENSE

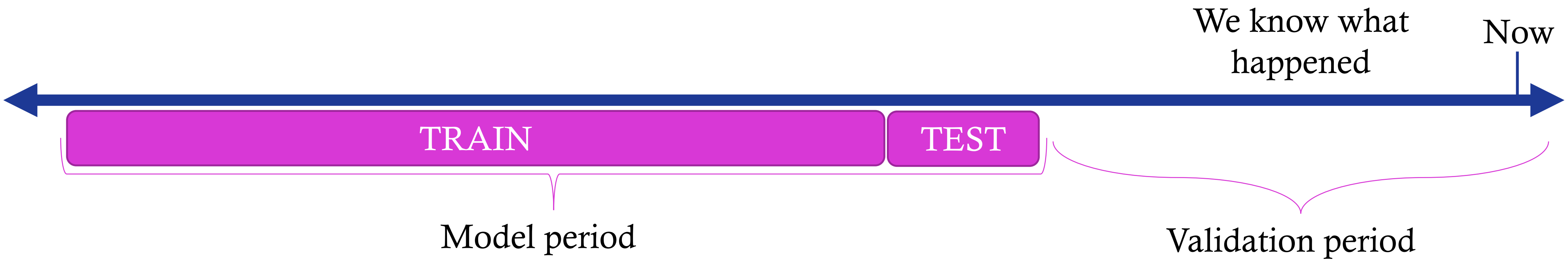
- Sometimes due to bias in the data some features might be extremely significant, but this cannot be explained business wise
- It's better to sacrifice accuracy if you can explain your results to your stakeholders



STEP 6: FOR RUNNING A ML PROJECT

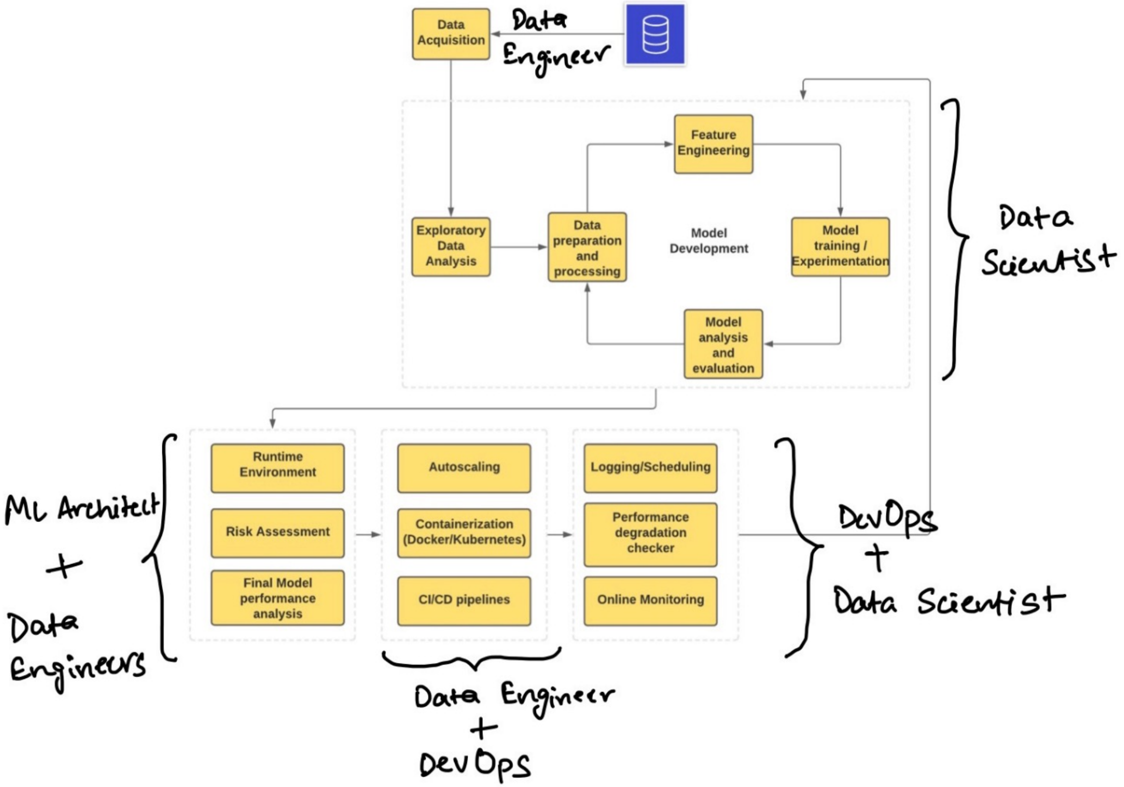
VALIDATE YOUR MODEL ON UNKNOWN DATA

- It's very common your model to perform perfectly on train-test dataset but when you tested on new data it won't work
- Keep some of your data to validate your model

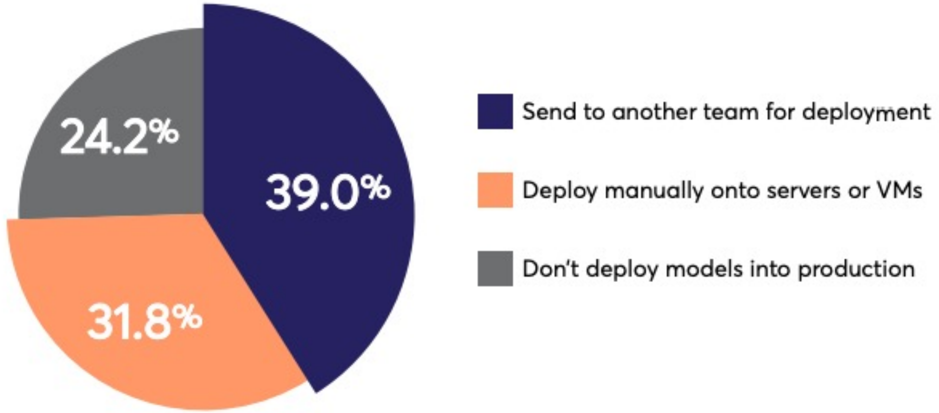


STEP 7: PUT IN PRODUCTION

REMEMBER THAT YOUR MODEL IS PART OF A SYSTEM



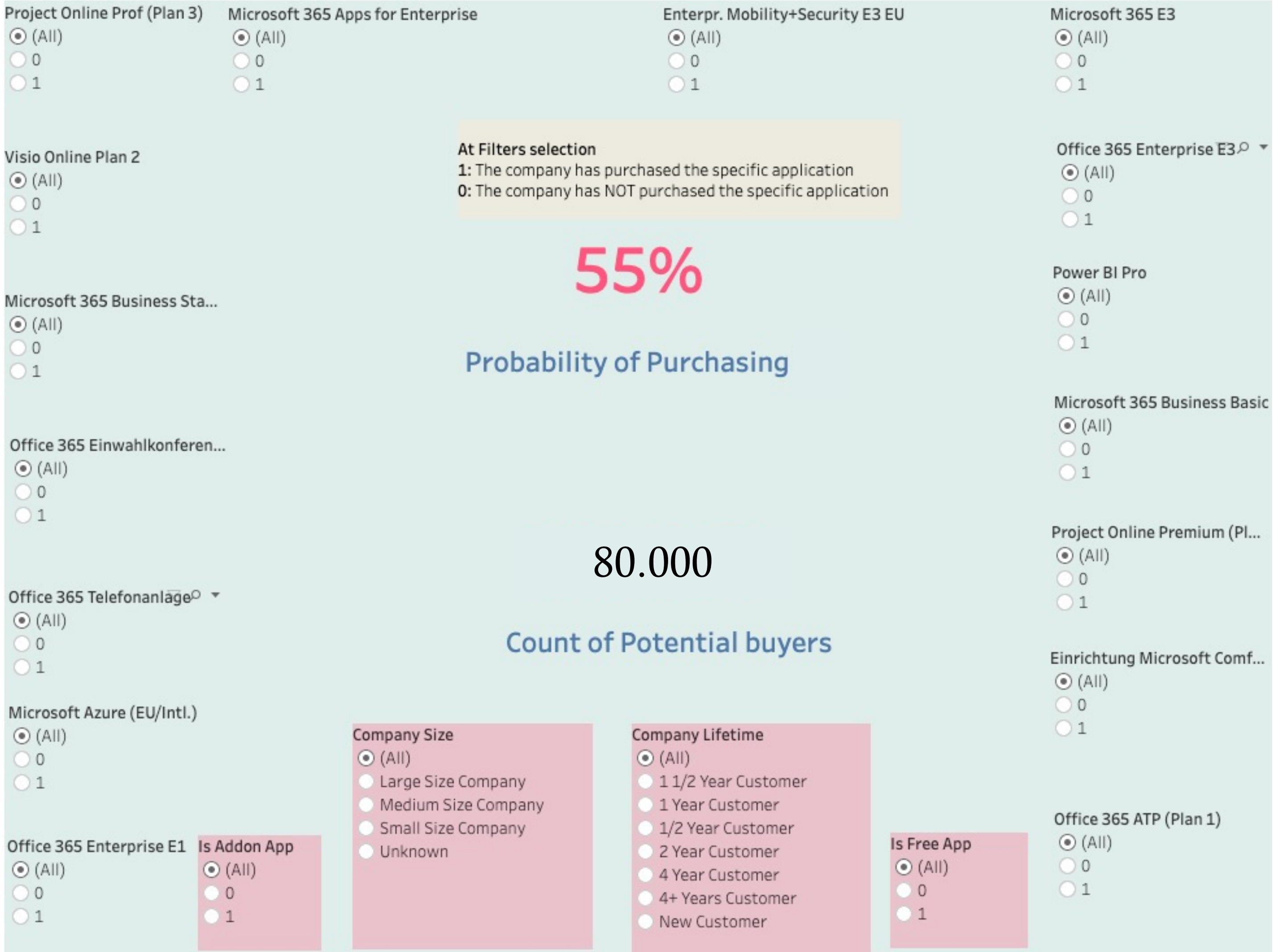
Nearly 40% of teams send models to another team for deployment into production (39%). The second most common response respondents gave is that their teams deploy them manually onto servers or VMs (31.8%). Another 24.2% stated that their ML and data science teams do not deploy models into production—18.6% of whom would like to.

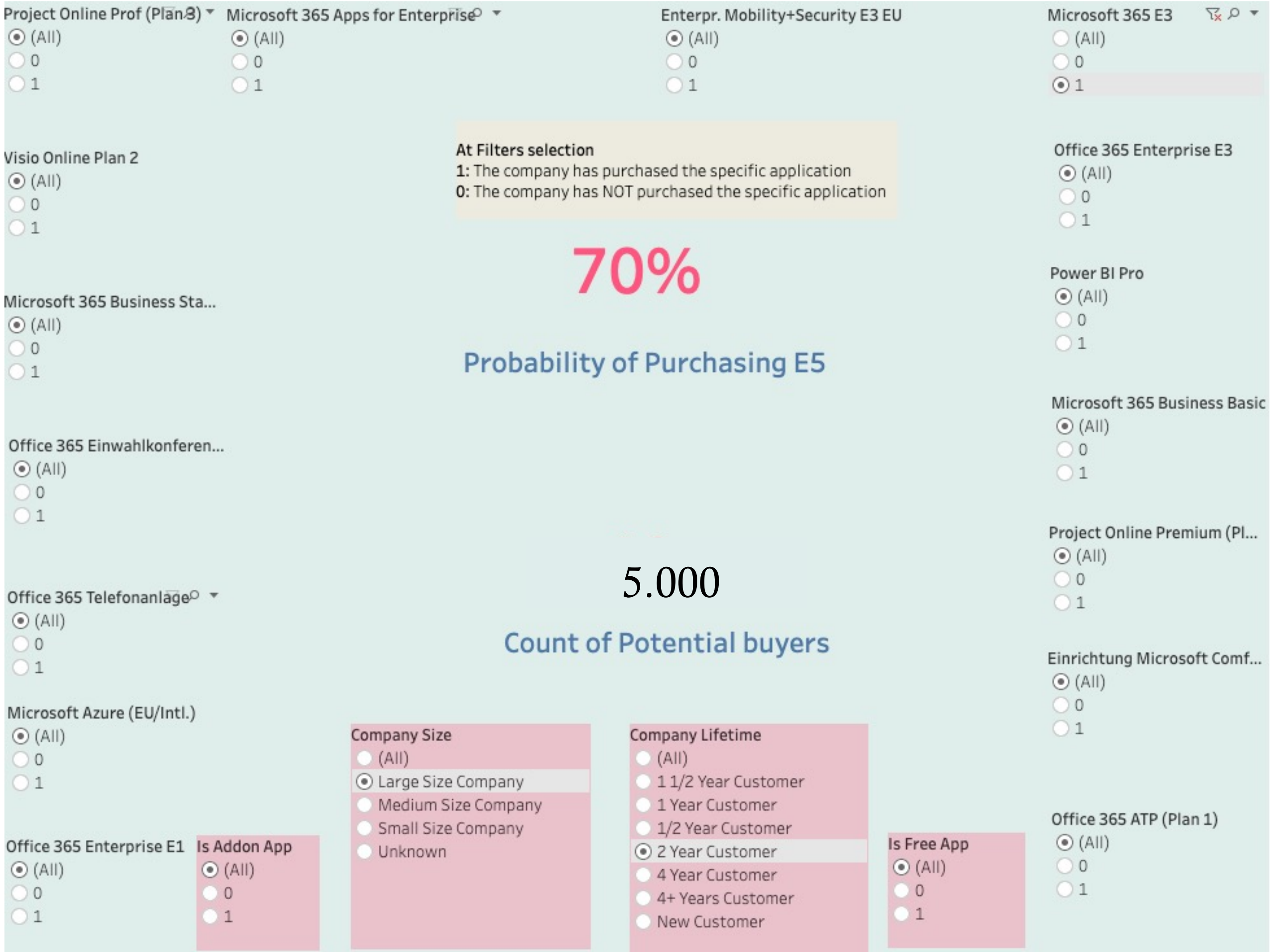


https://www.freecodecamp.org/news/what-is-mlops-machine-learning-operations-explained/
https://dotscience.com/assets/downloads/Dotscience_Survey-Report-2019.pdf

STEP 7: PUT IN PRODUCTION
REMEMBER THAT YOUR MODEL IS PART OF A SYSTEM

- User need to trust the model
- If possible, let the user interact with the model
- Give some info's why the model is predicting things
- Easy to use or well integrated into an existing business process


Project Online Prof (Plan 3)
(All)
0
1
Microsoft 365 Apps for Enterprise
(All)
0
1
Enterpr. Mobility+Security E3 EU
(All)
0
1
Microsoft 365 E3
(All)
0
1
Visio Online Plan 2
(All)
0
1
At Filters selection
1: The company has purchased the specific application
0: The company has NOT purchased the specific application
Office 365 Enterprise E3
(All)
0
1
55%
Probability of Purchasing
Power BI Pro
(All)
0
1
Microsoft 365 Business Sta...
(All)
0
1
Microsoft 365 Business Basic
(All)
0
1
Office 365 Einwahlkonferen...
(All)
0
1
Project Online Premium (Pl...
(All)
0
1
80.000
Count of Potential buyers
Office 365 Telefonanlage
(All)
0
1
Einrichtung Microsoft Comf...
(All)
0
1
Microsoft Azure (EU/Intl.)
(All)
0
1
Company Size
(All)
Large Size Company
Medium Size Company
Small Size Company
Unknown
Company Lifetime
(All)
1 1/2 Year Customer
1 Year Customer
1/2 Year Customer
2 Year Customer
4 Year Customer
4+ Years Customer
New Customer
Office 365 ATP (Plan 1)
(All)
0
1
Office 365 Enterprise E1
(All)
0
1
Is Addon App
(All)
0
1
Is Free App
(All)
0
1


Project Online Prof (Plan 3)
(All)
0
1
Microsoft 365 Apps for Enterprise
(All)
0
1
Enterpr. Mobility+Security E3 EU
(All)
0
1
Microsoft 365 E3
(All)
0
1
Visio Online Plan 2
(All)
0
1
At Filters selection
1: The company has purchased the specific application
0: The company has NOT purchased the specific application
Office 365 Enterprise E3
(All)
0
1
70%
Probability of Purchasing E5
Power BI Pro
(All)
0
1
Microsoft 365 Business Sta...
(All)
0
1
Microsoft 365 Business Basic
(All)
0
1
Office 365 Einwahlkonferen...
(All)
0
1
Project Online Premium (Pl...
(All)
0
1
Office 365 Telefonanlage
(All)
0
1
5.000
Count of Potential buyers
Einrichtung Microsoft Comf...
(All)
0
1
Microsoft Azure (EU/Intl.)
(All)
0
1
Company Size
(All)
Large Size Company
Medium Size Company
Small Size Company
Unknown
Company Lifetime
(All)
1 1/2 Year Customer
1 Year Customer
1/2 Year Customer
2 Year Customer
4 Year Customer
4+ Years Customer
New Customer
Office 365 Enterprise E1
(All)
0
1
Is Addon App
(All)
0
1
Is Free App
(All)
0
1
Office 365 ATP (Plan 1)
(All)
0
1

WE'RE HERE TO MAKE ML MODELS THAT MEET A BUSINESS NEED. MAKE SURE YOU UNDERSTAND THE NEED AND MEET IT AS EFFICIENTLY AS POSSIBLE.

Be open minded

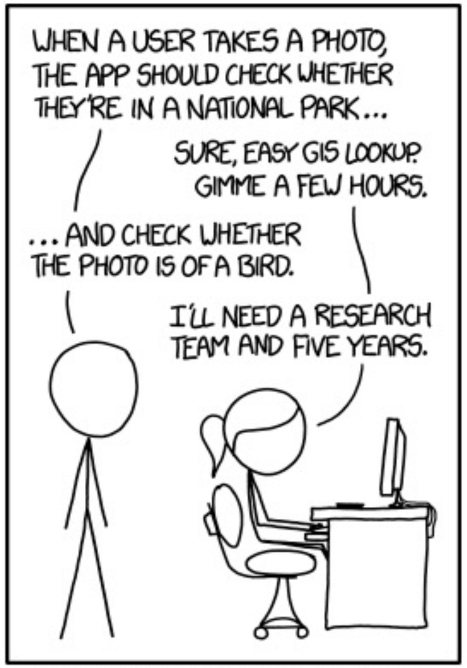
Be optimistic

Adaptability to learn the Business Context

Not lock into your tiny "ML-Room" - don't be isolated

A ML Model will never be 100% accurate (if it is .. it's wrong!)

Even ML can do great stuff (close to magic) it can't solve everything and will fail often.



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.