The role of Python in the data science lifecycle

Katinka Gereb, PhD

pyladies amsterdam

presents

Careers with python

Multi-series talks about career paths
as a Pythonista

24 Feb 2021

Nice to meet you

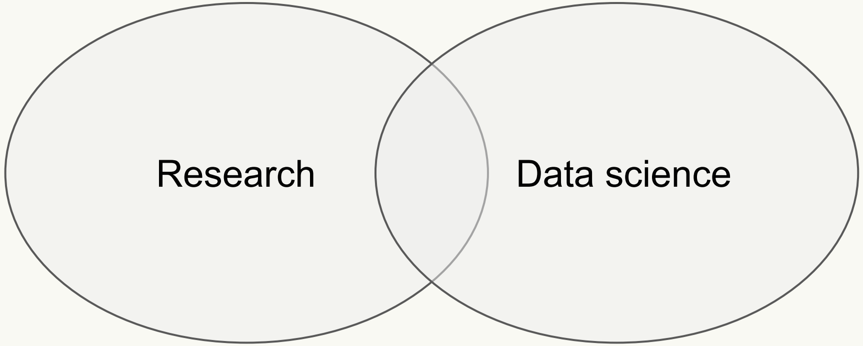
- Data scientist in Melbourne

Nice to meet you

- Data scientist in Melbourne

- PhD in astrophysics at the University of Groningen

How did I become a data scientist?



Research

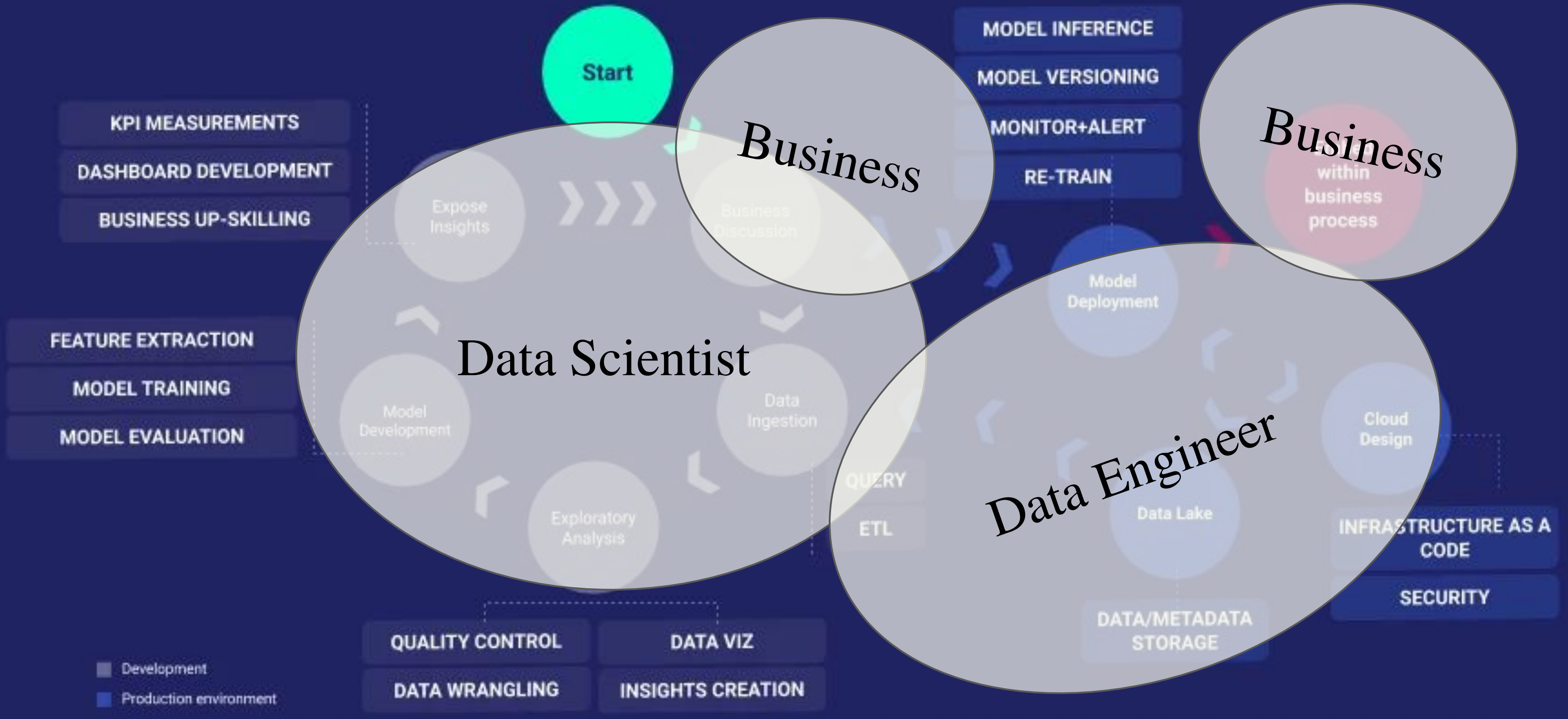
Data science

Why does a data scientist use Python?

Python can support the entire data science lifecycle!

- Data wrangling
- Statistical analysis
- Visualisations
- ML models
- Deployment

Data Science Cycle



Start

Business

Business

Data Scientist

Data Engineer

KPI MEASUREMENTS

DASHBOARD DEVELOPMENT

BUSINESS UP-SKILLING

Expose Insights

Business Discussion

Data Ingestion

Exploratory Analysis

Model Development

FEATURE EXTRACTION

MODEL TRAINING

MODEL EVALUATION

QUALITY CONTROL

DATA VIZ

DATA WRANGLING

INSIGHTS CREATION

MODEL INFERENCE

MODEL VERSIONING

MONITOR+ALERT

RE-TRAIN

Model Deployment

Embed within business process

Cloud Design

Data Lake

QUERY

ETL

DATA/METADATA STORAGE

INFRASTRUCTURE AS A CODE

SECURITY

▨ Development

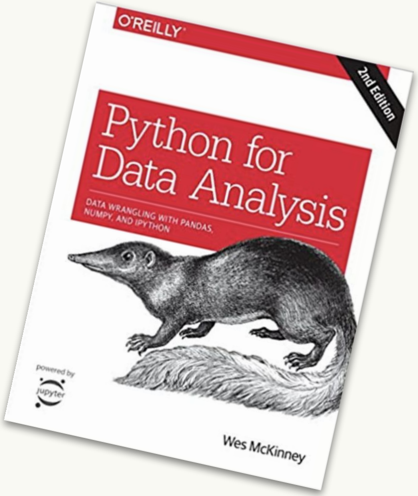
■ Production environment

Tools

Data cleaning and feature extraction

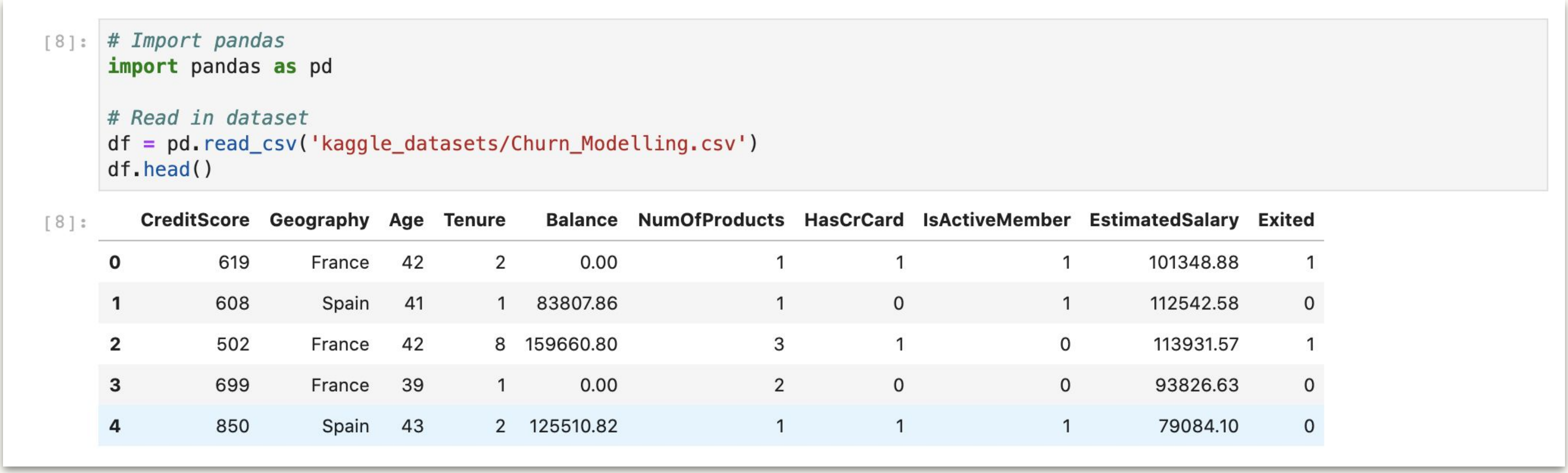
pandas

- DataFrame object for data manipulation
- Reading and writing
- Label-based computations
- Missing data, reshaping
- Group by, merge, concat
- Time-series functionality



Data cleaning and feature extraction

pandas



```
[8]: # Import pandas
     import pandas as pd

     # Read in dataset
     df = pd.read_csv('kaggle_datasets/Churn_Modelling.csv')
     df.head()
```

```
[8]:    CreditScore  Geography  Age  Tenure    Balance  NumOfProducts  HasCrCard  IsActiveMember  EstimatedSalary  Exited
     0          619     France   42       2       0.00              1          1               1        101348.88       1
     1          608      Spain   41       1   83807.86              1          0               1        112542.58       0
     2          502     France   42       8  159660.80              3          1               0        113931.57       1
     3          699     France   39       1       0.00              2          0               0         93826.63       0
     4          850      Spain   43       2  125510.82              1          1               1         79084.10       0
```

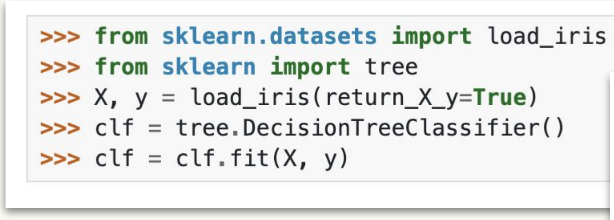
Machine learning

- Simple and efficient tools for predictive data analysis
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable

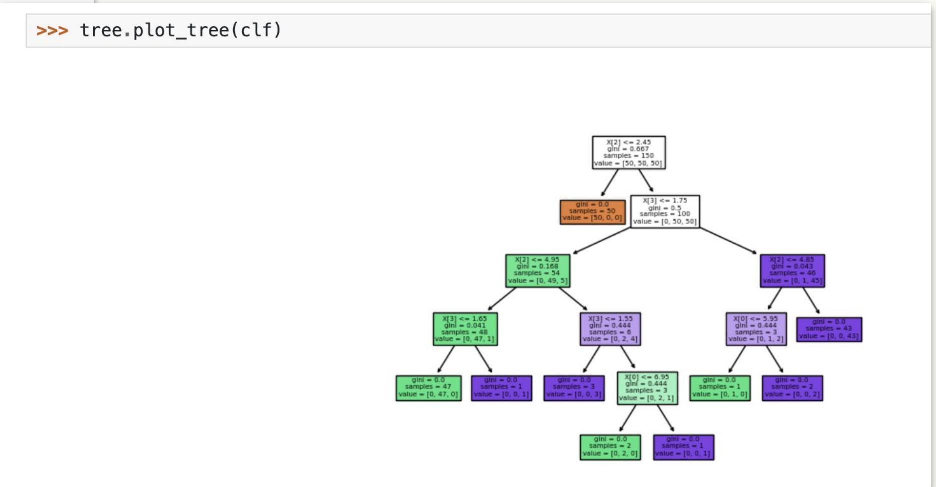
Machine learning

scikits learn

machine learning in Python



```
>>> from sklearn.datasets import load_iris
>>> from sklearn import tree
>>> X, y = load_iris(return_X_y=True)
>>> clf = tree.DecisionTreeClassifier()
>>> clf = clf.fit(X, y)
```



```
>>> tree.plot_tree(clf)
```

X[2] <= 2.45 gini = 0.667 samples = 150 value = [50, 50, 50]
gini = 0.0 samples = 50 value = [50, 0, 0]
X[3] <= 1.75 gini = 0.5 samples = 100 value = [0, 50, 50]
X[2] <= 4.95 gini = 0.168 samples = 54 value = [0, 49, 5]
X[2] <= 4.85 gini = 0.043 samples = 46 value = [0, 1, 45]
X[3] <= 1.65 gini = 0.041 samples = 48 value = [0, 47, 1]
X[3] <= 1.55 gini = 0.444 samples = 6 value = [0, 2, 4]
X[0] <= 5.95 gini = 0.444 samples = 3 value = [0, 1, 2]
gini = 0.0 samples = 43 value = [0, 0, 43]
gini = 0.0 samples = 47 value = [0, 47, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.0 samples = 3 value = [0, 0, 3]
X[0] <= 6.95 gini = 0.444 samples = 3 value = [0, 2, 1]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.0 samples = 2 value = [0, 0, 2]
gini = 0.0 samples = 2 value = [0, 2, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]

Data visualisation

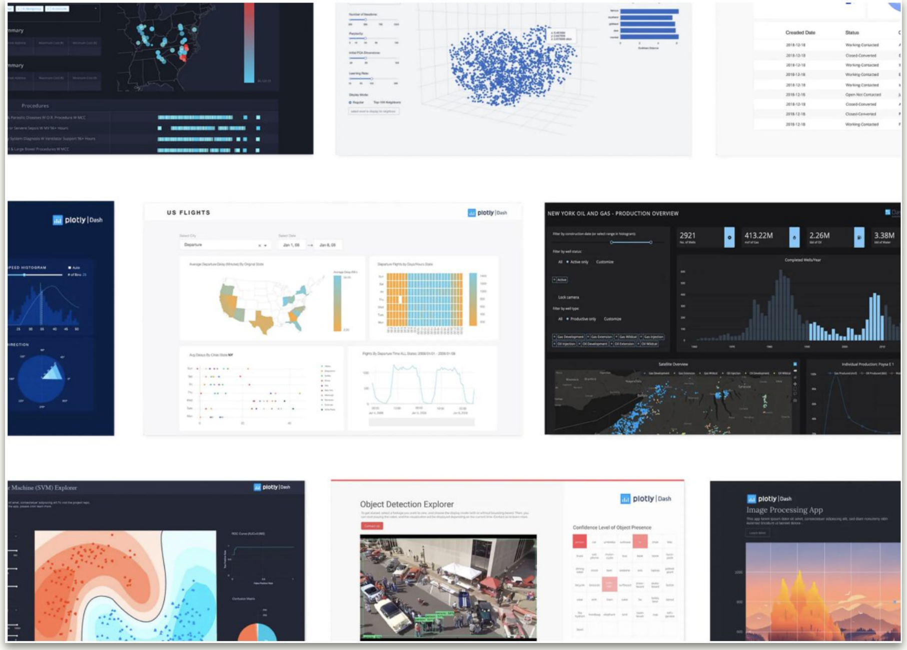
plotly

- Interactive visualisations

Web apps

plotly

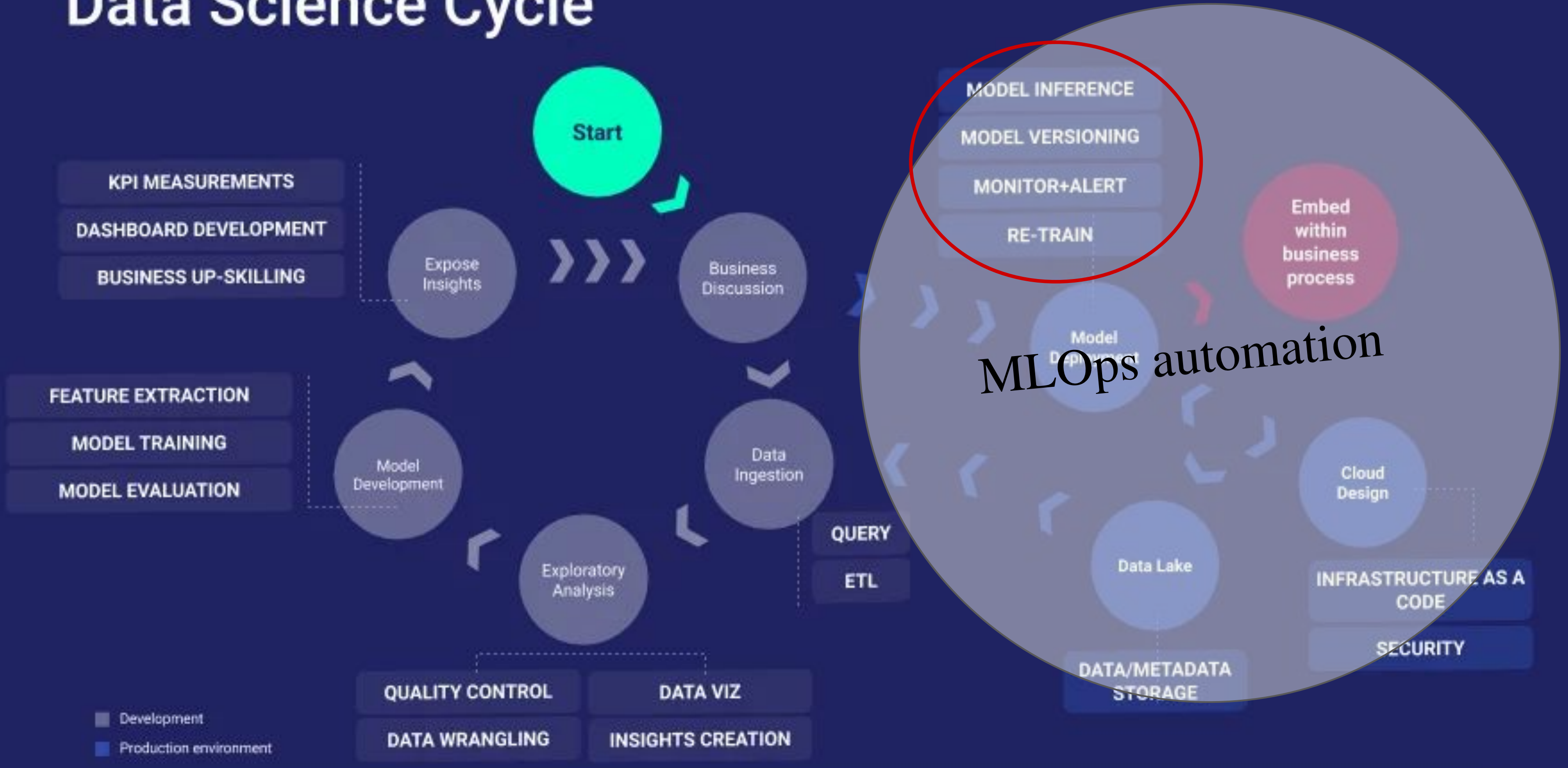
- Build and deploy apps



US FLIGHTS
plotly | Dash
NEW YORK OIL AND GAS - PRODUCTION OVERVIEW
2921
413.22M
2.26M
3.38M
Machine (SVM) Explorer
Object Detection Explorer
Confidence Level of Object Presence
Image Processing App

Towards machine learning automation

Data Science Cycle



Start

Business Discussion

Data Ingestion

QUERY

ETL

Exploratory Analysis

QUALITY CONTROL

DATA VIZ

DATA WRANGLING

INSIGHTS CREATION

Model Development

FEATURE EXTRACTION

MODEL TRAINING

MODEL EVALUATION

Expose Insights

KPI MEASUREMENTS

DASHBOARD DEVELOPMENT

BUSINESS UP-SKILLING

MODEL INFERENCE

MODEL VERSIONING

MONITOR+ALERT

RE-TRAIN

Model Deployment

Embed within business process

Cloud Design

INFRASTRUCTURE AS A CODE

SECURITY

Data Lake

DATA/METADATA STORAGE

MLOps automation

▨ Development

▨ Production environment

MLOps

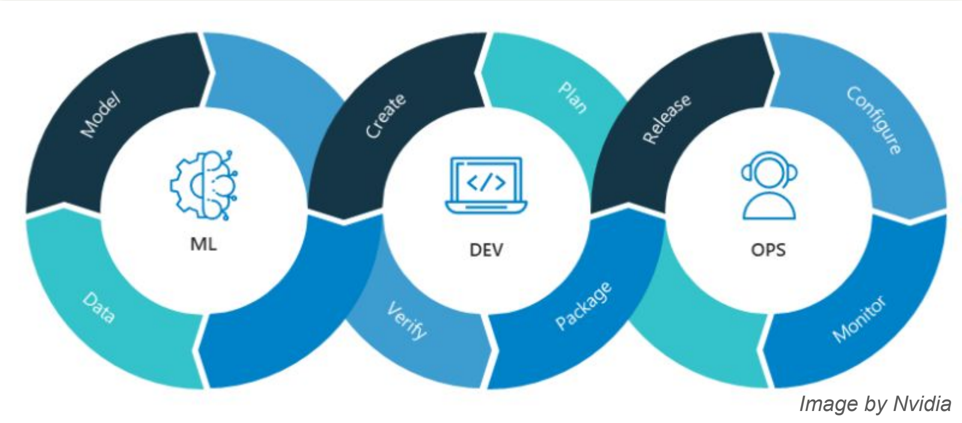


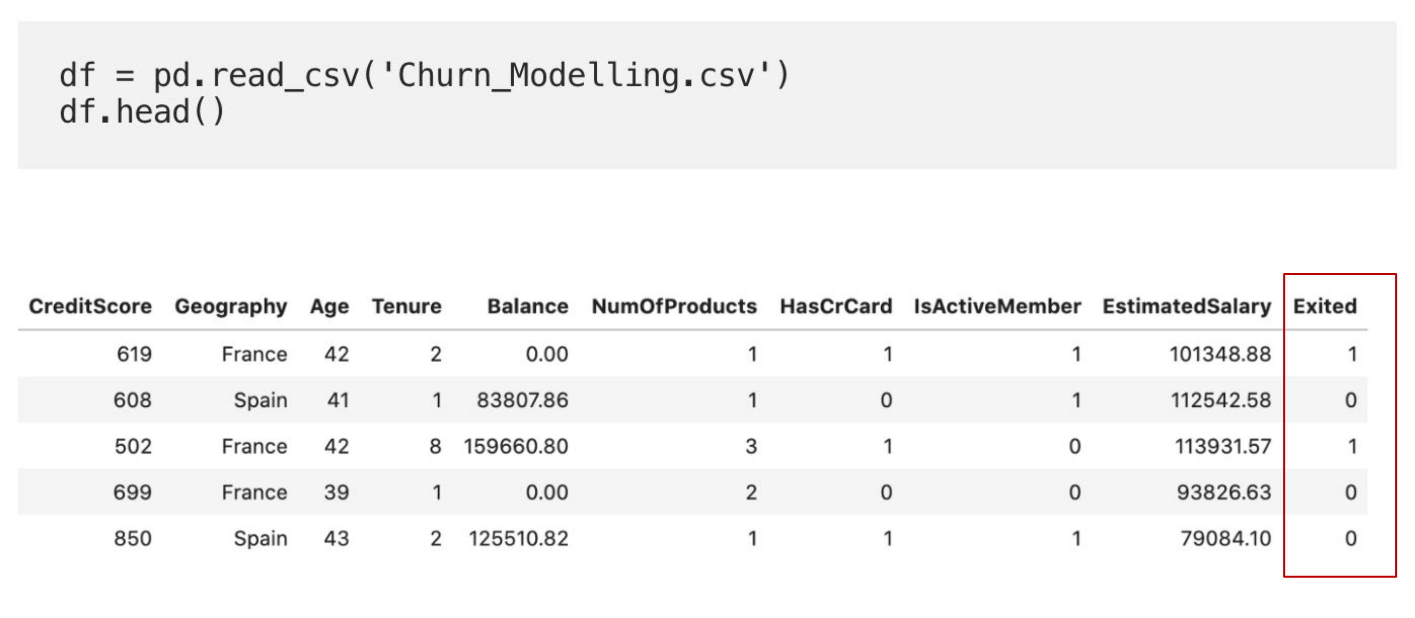
Image by Nvidia

- Trackable
- Reproducible
- Self-sustaining
- Automated

MLOps in the cloud

Use cases

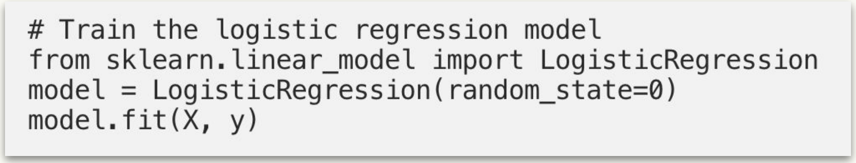
Churn Prediction



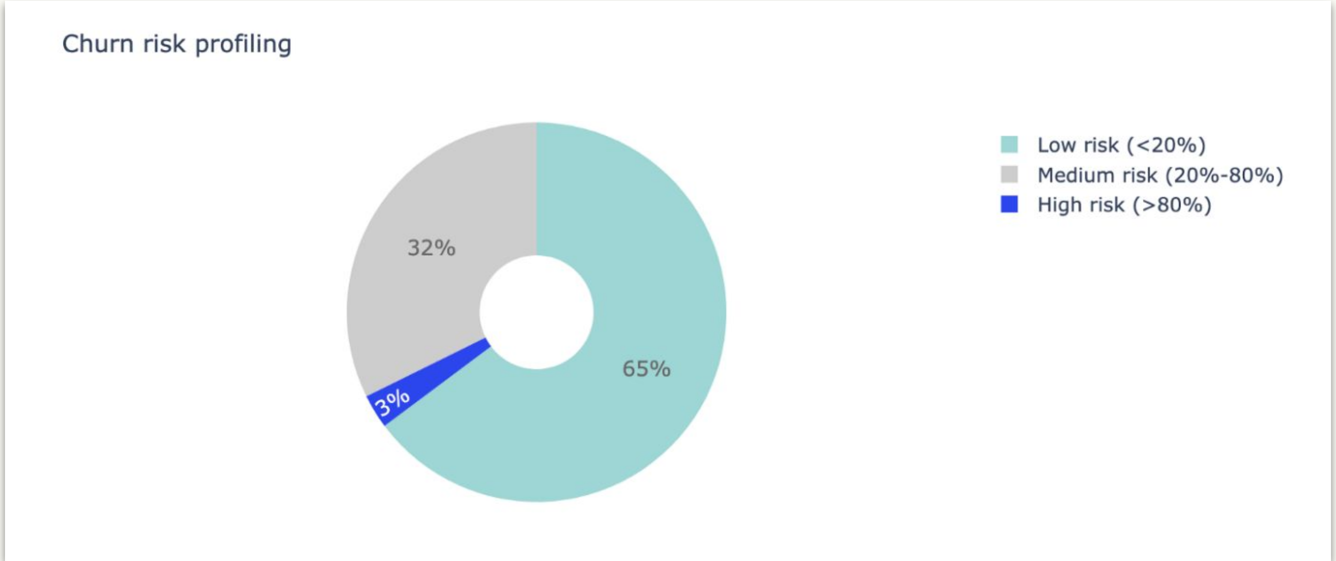
```
df = pd.read_csv('Churn_Modelling.csv')
df.head()
```

CreditScore	Geography	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
619	France	42	2	0.00	1	1	1	101348.88	1
608	Spain	41	1	83807.86	1	0	1	112542.58	0
502	France	42	8	159660.80	3	1	0	113931.57	1
699	France	39	1	0.00	2	0	0	93826.63	0
850	Spain	43	2	125510.82	1	1	1	79084.10	0

Churn Prediction



```
# Train the logistic regression model
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=0)
model.fit(X, y)
```

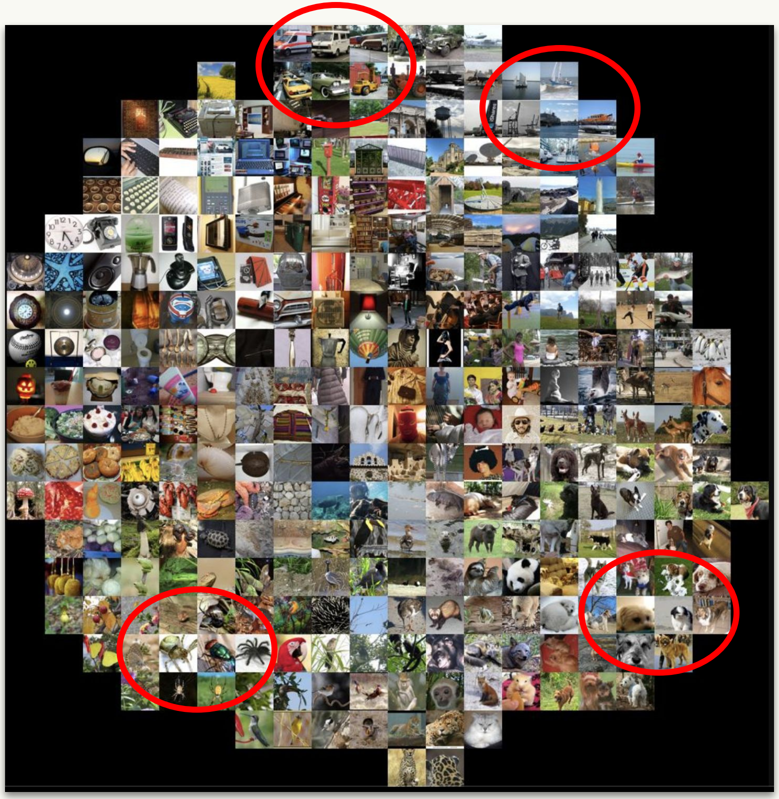


Risk profiling for customer churn analysis

Katinka Gereb Aug 1, 2020 · 6 min read

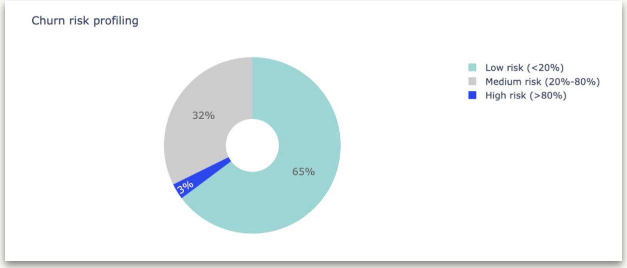
@katinka-gereb

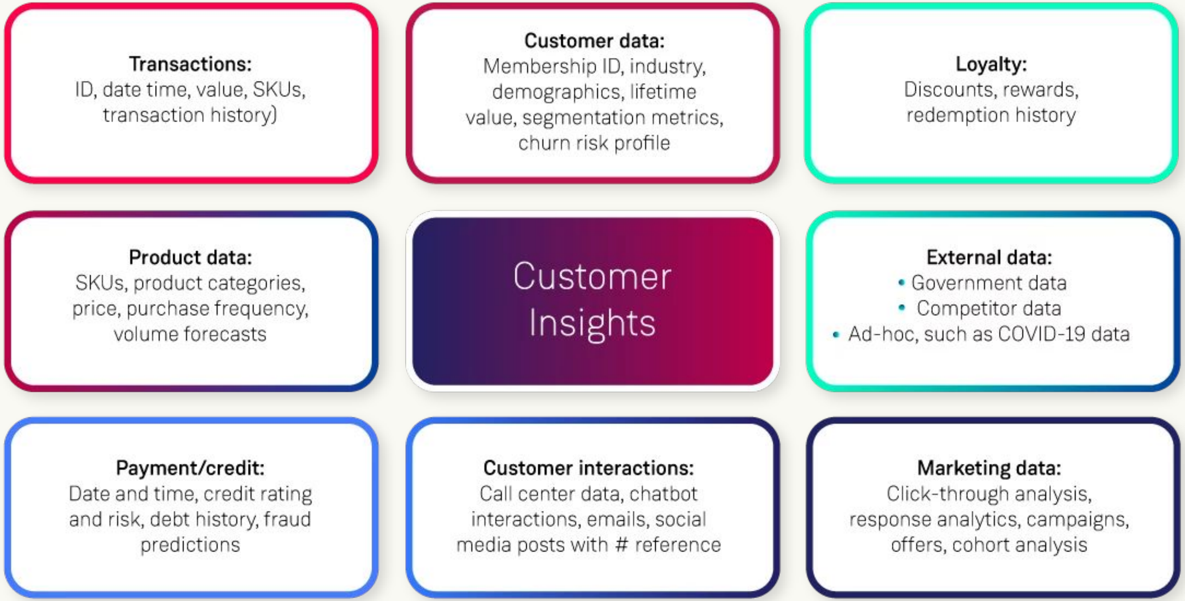
Clustering



https://cs.stanford.edu/people/karpathy/cnnembed/

Customer 360





Transactions:
ID, date time, value, SKUs, transaction history)

Customer data:
Membership ID, industry, demographics, lifetime value, segmentation metrics, churn risk profile

Loyalty:
Discounts, rewards, redemption history

Product data:
SKUs, product categories, price, purchase frequency, volume forecasts

Customer Insights

External data:
- Government data
- Competitor data
- Ad-hoc, such as COVID-19 data

Payment/credit:
Date and time, credit rating and risk, debt history, fraud predictions

Customer interactions:
Call center data, chatbot interactions, emails, social media posts with # reference

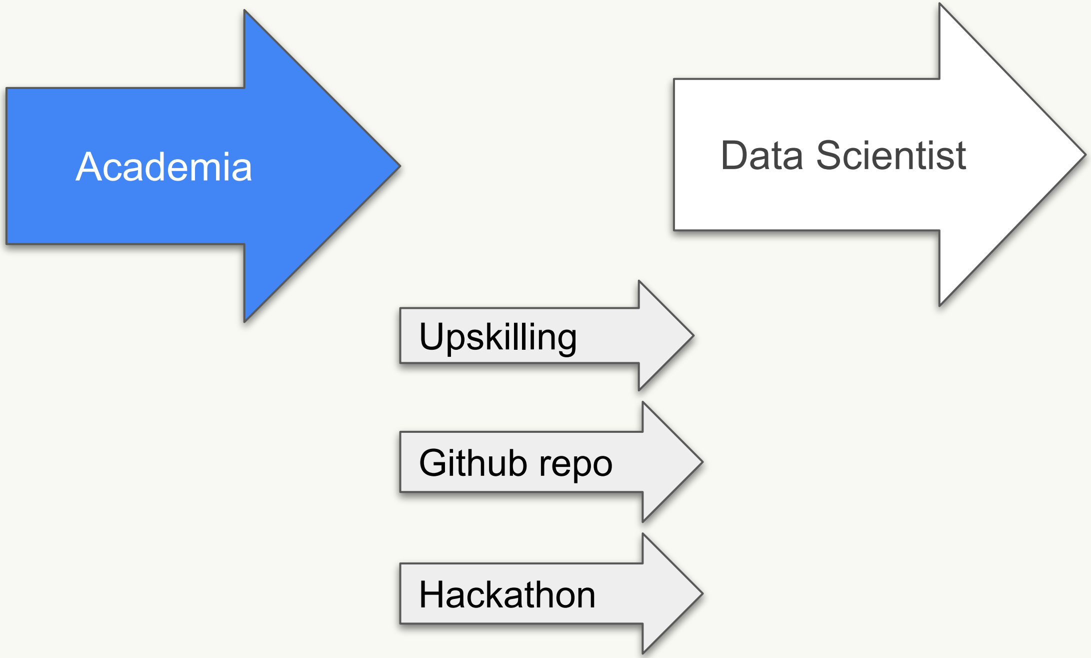
Marketing data:
Click-through analysis, response analytics, campaigns, offers, cohort analysis

Advancing your career

Academia

Data Scientist

Advancing your career



Academia

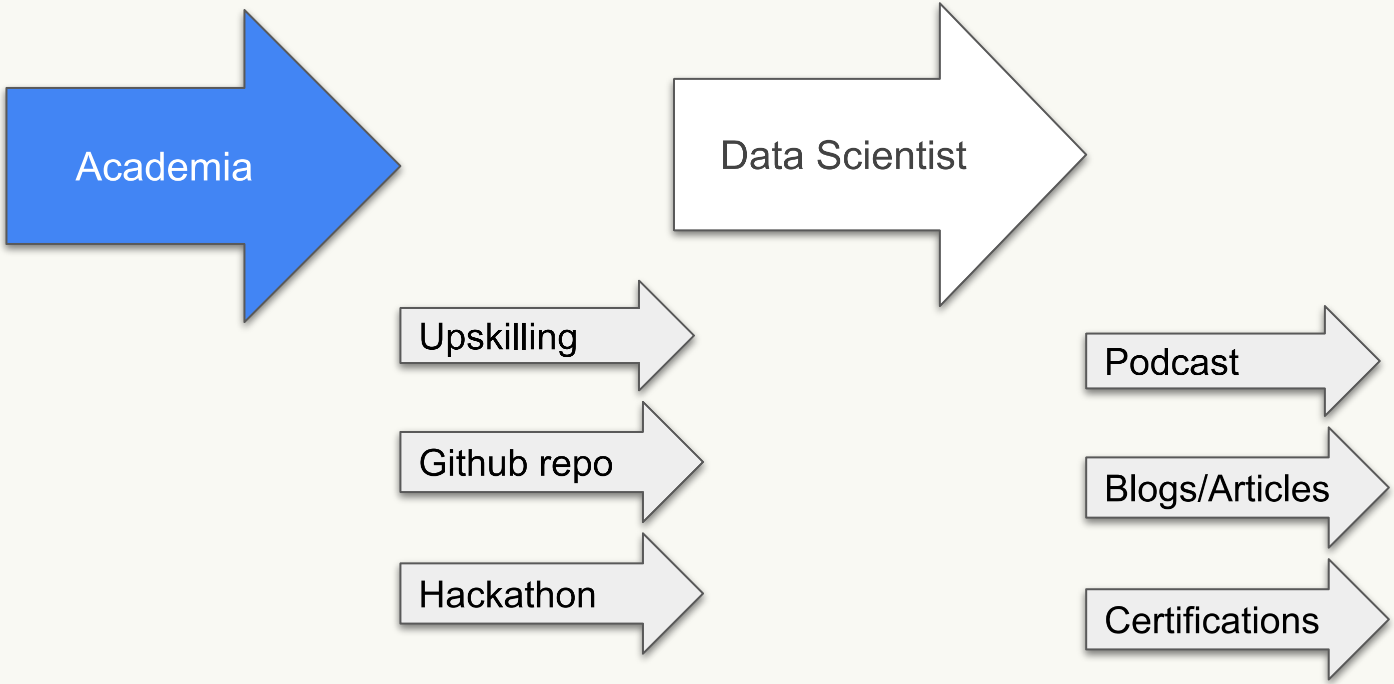
Data Scientist

Upskilling

Github repo

Hackathon

Advancing your career



Academia

Data Scientist

Upskilling

Github repo

Hackathon

Podcast

Blogs/Articles

Certifications

@katinka-gereb

Thanks for listening!

Nova Tech Podcast

LEARN, ADAPT, SUCCEED

Nova Tech Podcast

https://novatech.buzzsprout.com/