Project Title	End to end data science
Tools	Python
Technologies	Data science
Project Difficulties level	intermediate

About dataset

The end to end data science pipeline is the complete processof solving a problem solving from collecting and preparing data to model, evaluate.

End to end analytics is a tool to tracking customer data,sales,advertising compaigons ,and more.It helps identify popular products,effevtive marketing channels and lucrative investments, Financial institutions .It is usefull for banks, insurance company and so on.

Data is encrypted on the sender's device and is only ever decrypted on the recipients device - never in the cloud - because only the sender and recipient possess the keys to encrypt and decrypt the message.As a result, attackers watching internet traffic or breaching a server cannot access the data.

An End-to-End Data Science Project

Session 1
Preparation
10.04.2022

Start with the business problem, find data source, preprocess data, set up team

Session 2
Analytics
17.04.2022

Analyze and understand your data. Gain insights and prepare for the predictive modeling

Session 3
Machine learning
x.05.2022

Build and evaluate prediction model(s), use Mlflowto keep track of the various experiments

Session 4
Production
x.05.2022

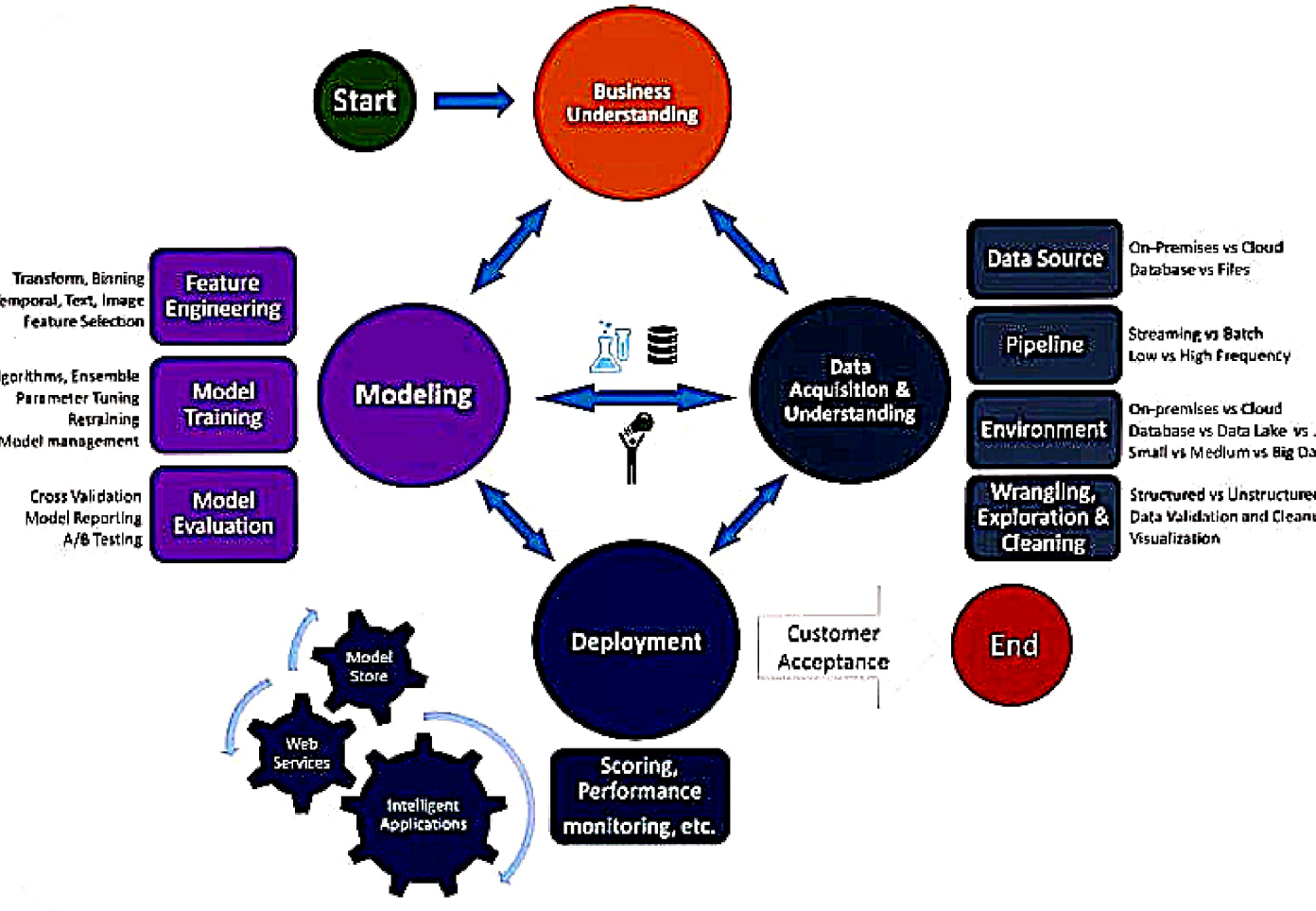
Create prediction functions and production class, develop an API, create a dashboard that the user will access and

What you will do:

Session 1: Recap

Session 1: Main points

Build a business case Find suitable data sources Verify legal rights Track your project via Git Explore and preprocess data Collaborate with your team using Kanban


Start
Business Understanding
Feature Engineering
Transform, Binning
emporal, Text, Image
Feature Selection
Modeling
Model Training
gorithms, Ensemble
Parameter Tuning
Retraining
Model management
Model Evaluation
Cross Validation
Model Reporting
A/B Testing
Data Acquisition & Understanding
Data Source
On-Premises vs Cloud
Database vs Files
Pipeline
Streaming vs Batch
Low vs High Frequency
Environment
On-premises vs Cloud
Database vs Data Lake vs
Small vs Medium vs Big Da
Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleans
Visualization
Deployment
Customer Acceptance
End
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

Session 1: Main points

- *Form your team and create your Kanban board*
- *Create your **project directory** and track in a new **GitHub** repo*
- *Preprocess your raw data and export it to a pickle file*
- *Complete your **descriptive analytics** part – understand your data and get insights to be used in the modelling*

Session 2: Descriptive Analytics

Part 1.
Insights

“Asking the right question is half of the answer”

It's your turn:
What are the descriptive questions that you will answer ?

Think about what you want to do before you start doing it. Keep the original goal in mind

My analytics question

General:
- Total number of answers
- Geographical distributions
- Missing answers

Skills:
Frequency of each skill
How are the skills correlated with each others

Jobs:
Frequency of each job
How are the jobs correlated with each others

Relation:
How are the skills correlated with the jobs
- What is the specificity of each skill to a job

Levels of descriptive analytics

1. *Stats or summary tables*
2. *Visualizations*
3. *Unsupervised learning (e.g. clustering)*

Part 2. Unsupervised to Supervised

Unsupervised to Supervised

T-SNE

Stands for t-distributed stochastic neighbor embedding.
Nonlinear dimensionality reduction technique

Agglomerative Clustering

Recursively merges the pair of clusters that minimally increases a given linkage distance

Silhouette metric

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

Part 3.
Data manipulation

Data selection

Responses: Select responses within reasonable ranges

Classes:
- Drop non-relevant classes (e.g.Senior executive)
- Merge close classes (e.g.Scientist & Researcher)
- Split vague classes (e.g.Backend developer)

Features:
- Create new features (e.g.Skills groups)
- Drop irrelevant features (e.g.Platforms)

- *Complete and enhance your descriptive analytics pipeline*
- *Start with the predictive analytics (X: Skills, Y: Jobs)*

An End-to-End Data Science Project

Workshop overview:

Session 1
Preparation
10.04.2022

Start with the business problem, find data source, preprocess data, set up team

Session 2
Analytics
17.04.2022

Analyze and understand your data. Gain insights and prepare for the predictive modeling

Session 3
Machine learning
x.05.2022

Build and evaluate prediction model(s), use Mlflowto keep track of the various experiments

Session 4
Production
x.05.2022

Create prediction functions and production class, develop an API, create a dashboard that the user will access and

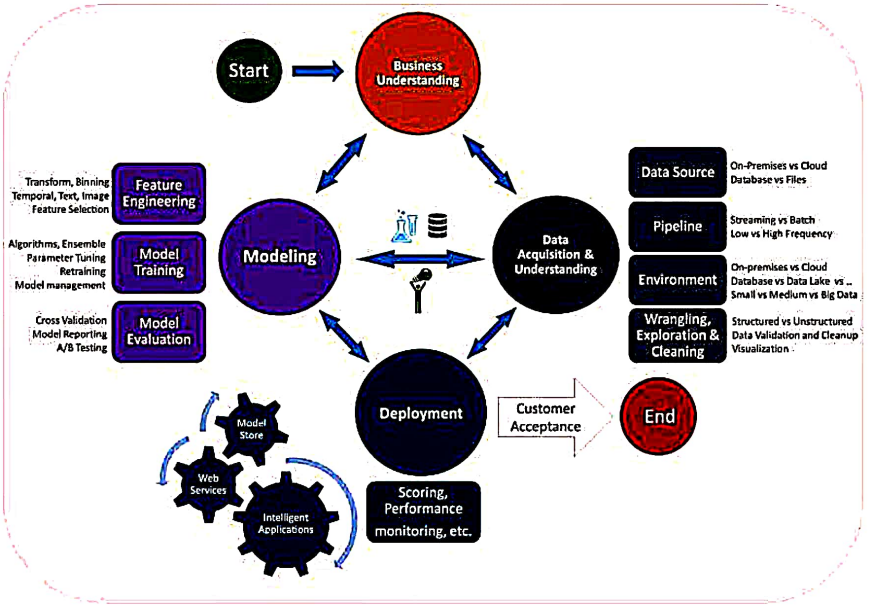
What you will do:

Session 1: Recap

Session 1: Main points

Build a business case Find suitable data sources Verify legal rights Track your project via Git Explore and preprocess data Collaborate with your team using Kanban

Session 1: Main points



Session 1: Main points

- *Form your **team**and create your **Kanban**board*
- *Create your **project directory**and track in a new **GitHub** repo*
- ***Preprocess**your raw data and export it to a pickle file*
- *Complete your **descriptive analytics** part – understand your data and get insights to be used in the modelling*

Session 2: Descriptive Analytics

Part 1.
Insights

“Asking the right question is half of the answer”

It's your turn:
What are the descriptive questions that you will answer ?

Think about what you want to do before you start doing it. Keep the original goal in mind

My analytics question

General:
- Total number of answers
- Geographical distributions
- Missing answers

Skills:
- Frequency of each skill
- How are the skills correlated with each others

Jobs:
- Frequency of each job
- How are the jobs correlated with each others

Relation:
- How are the skills correlated with the jobs
- What is the specificity of each skill to a job

Levels of descriptive analytics

1. *Stats or summary tables*
2. *Visualizations*
3. *Unsupervised learning (e.g. clustering)*

Part 2. Unsupervised to Supervised

Unsupervised to Supervised

T-SNE

Stands for t-distributed stochastic neighbor embedding.
Nonlinear dimensionality reduction technique

Agglomerative Clustering

Recursively merges the pair of clusters that minimally increases a given linkage distance

Silhouette metric

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)