ML teams and projects management: potential for cost optimization

Mikhail Rozhkov

ml-repa.ru

Outline

1. ML project workflow
2. The role of experiments in ML Workflow
3. ML Experiments issues
4. Story of one attempt

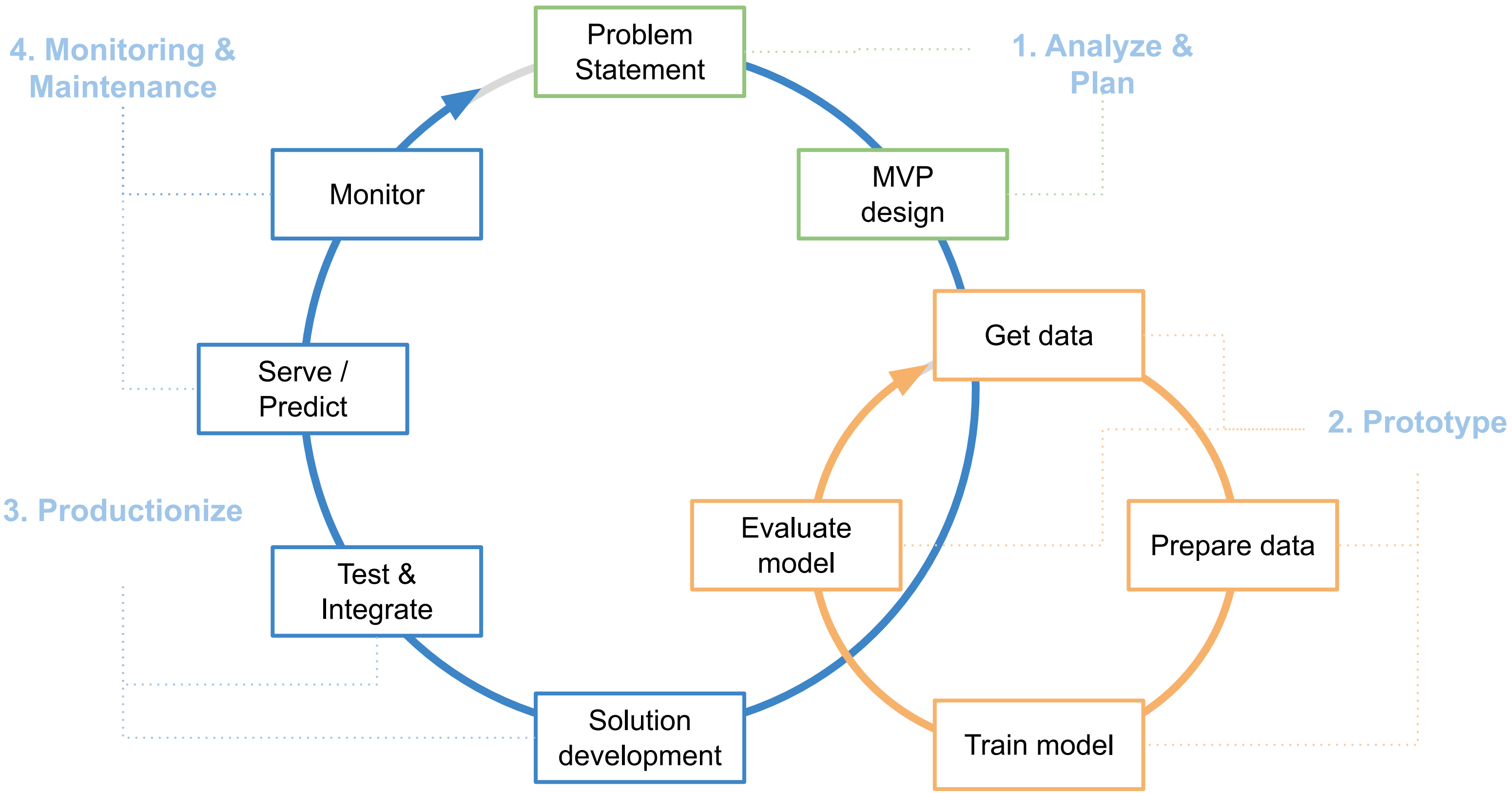
Common DS/ML issues

- Fragmented ML practices
- Difficult sharing & collaboration
- Inefficiency & Work duplication
- Updates are slow
- Pipelines not reliable or reproducible
- Scalability performance
- Data quality issues
- Model & features monitor and discoverability

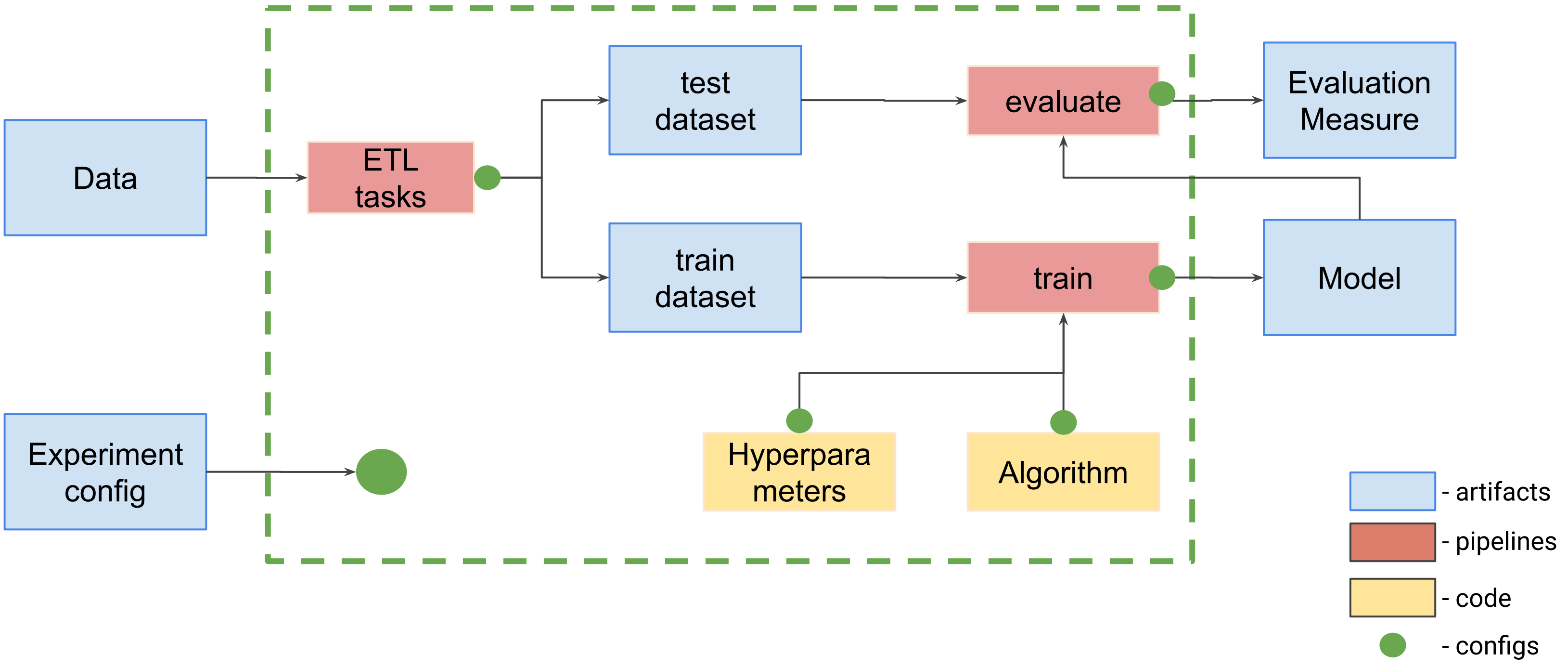
Why?

1. Different from IT projects
2. Longer dev cycle
3. Experiments driven
4. Not easy to test and validate

ML project workflow is experiments driven



Experiment = code + dataset + outputs



ML project requires more factors to take into account

	Software	ML
Architecture design	tasks, UI/UX integrations	+ nature and quality of data
Quality measures	working code	+ model quality metrics + performance in production
Version control	code environment	+ pipelines + datasets + models & artifacts
Testing	code	+ data and features + model development methods + ML infrastructure + ML systems

Inspired by Dmitry Petrov, Ivan Shcheklein. Open source tools for machine learning model and dataset versioning.

Where is value?

- Goal - **get economic value**
- Objectives:
  - **increase sales**
  - **optimize costs**
- The only way - **integrate into business process**

Fast experimentation helps to Fail or Success faster

1. Satisfy customer
2. Early fail
3. Fail safe
4. Team = Business + DS/ML
5. Constant changing requirements
6. Frequent team meetings/statuses
7. Frequent code updates
8. Measure of progress=working code
9. Technical excellence and good design
10. Reproducibility

Management & Organizational practices

Software Engineering practices

Inspired by: Andrew Kelleher, Adam Kelleher. Machine Learning in Production: Developing and Optimizing Data Science Workflows and Applications. 2019

ML Experiments issues

What are issues in ML Experiments?

Issues:

1. Difficult collaboration & Work duplication

Sources:

- all code in Jupyter Notebooks
- difficult to share / re-use
- copy-paste development
- no versioning
- no shared Storage
- no Task Tracking

Lack of software engineering best practices

What are issues in ML Experiments?

Issues:

1. Difficult collaboration & Work duplication
2. Models and Artifacts versioning

Sources:

Lack of software engineering best practices

- Manual naming / versioning
- No special tools
- No shared storage for artifacts
- No model registry

No model & artifacts version control

What are issues in ML Experiments?

Issues:

1. Difficult collaboration & Work duplication
2. Models and Artifacts versioning
3. Pipelines reproducibility

Sources:

Lack of software engineering best practices

No model & artifacts version control

- Not automated pipelines
- No control of run params
- Environment dependencies control

Not reproducible pipelines

What are issues in ML Experiments?

Issues:

1. Difficult collaboration & Work duplication
2. Models and Artifacts versioning
3. Pipelines reproducibility
4. Experiments benchmarking

Sources:

Lack of software engineering best practices

No model & artifacts version control

Not reproducible pipelines

- No experiment management tools
- No experimentation culture

No experiment management

What are issues in ML Experiments?

Issues:

1. Difficult collaboration & Work duplication
2. Models and Artifacts versioning
3. Pipelines reproducibility
4. Experiments benchmarking
5. Updates are slow

Sources:

Lack of software engineering best practices

No model & artifacts version control

Not reproducible pipelines

No experiment management

Story of one attempt

not official point of view

lasts 10 months

in progress...

Start position (WAS IS)

1. No a Department or Head responsible for ML development
2. Few autonomous teams (~30 DS in total)
3. Different engineering background and tasks
4. No cross-team projects
5. No common standards on how to do things
6. Almost all job is done in Jupyter Notebooks
7. Few models in production (manually)
8. Complicated enterprise IT infrastructure

Common problems:

- Work duplication
- No version control
- Difficult to reproduce pipelines
- No project documentation
- Updates are slow

Task: make it in right way

1. Try / select appropriate tools
2. Apply in your own project
3. Convince other DS to try / use
4. Share knowledge & help
5. Estimate economic value of changes
6. Plan / implement changes

mlflow

DVC

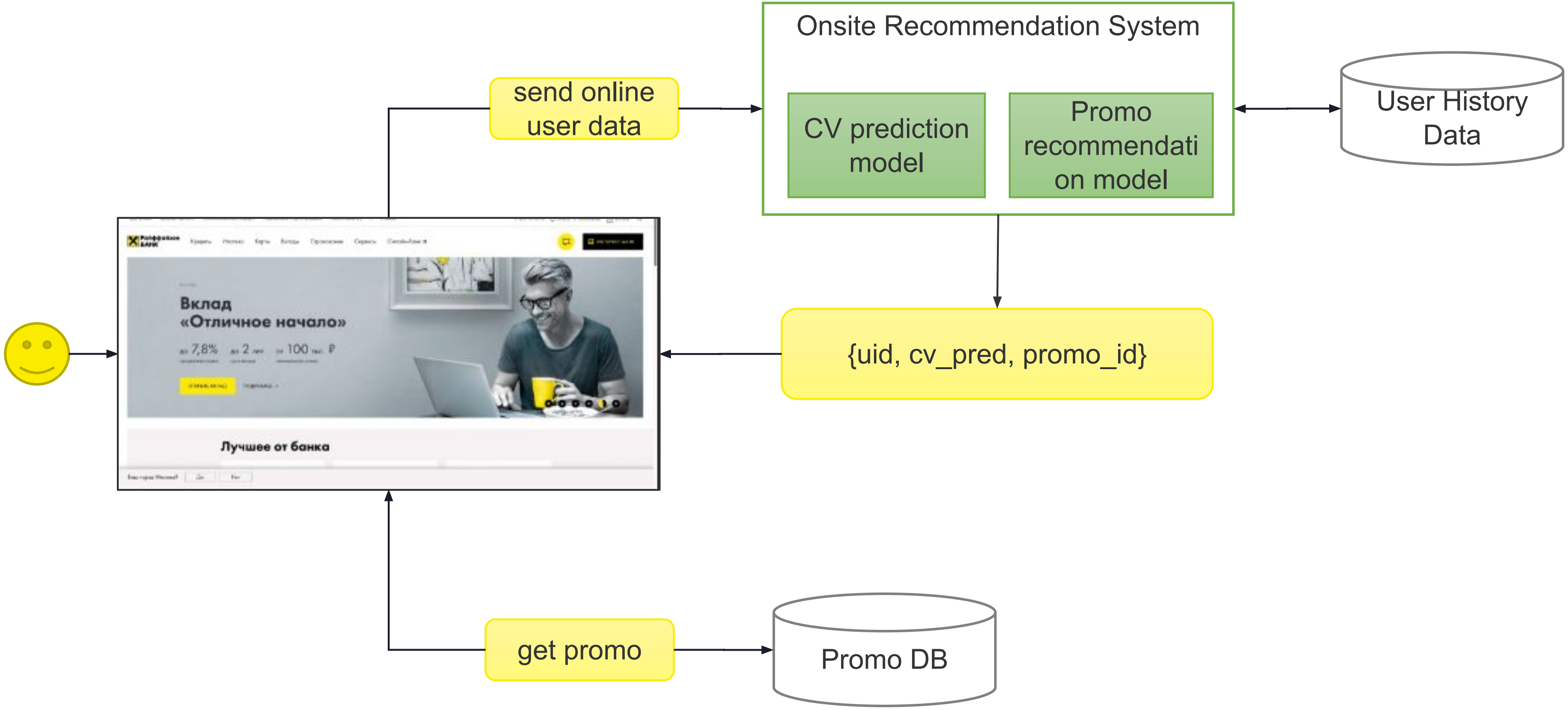
Bitbucket

JIRA

Confluence

Atlassian Bamboo

Use case: Predict a new user propensity to apply for Credit Card on Landing Page



Documentation and tracking project statuses

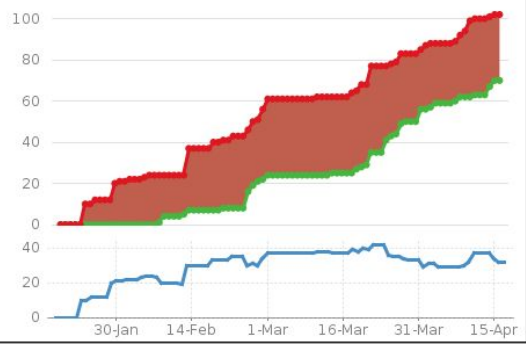
Project Profile

Topic	
Business Target	
Domain	
Type	
Project ID	
Status	
Customer	
Department - Initiator	
Department - Recepient	
Directorate	
Area	
B-1	
Department in charge	
Team Lead	
Analysts	
Data Scientists	

Project Issues

Key	Summary	T	Created	Updated	Status
BDAA-67	1. Submit Business Request (hypotheses)	☑	Jan 23, 2019	Feb 12, 2019	RESOLVED
BDAA-68	2. Pre-study	☑	Jan 23, 2019	Feb 25, 2019	RESOLVED
BDAA-69	3. Prepare data	☑	Jan 23, 2019	Feb 13, 2019	IN PROGRESS
BDAA-70	4. Method & Research Design	☑	Jan 23, 2019	Apr 01, 2019	RESOLVED
BDAA-71	5. R&D	☑	Jan 23, 2019	Feb 13, 2019	IN PROGRESS
BDAA-72	6. Testing & validation	☑	Jan 23, 2019	Apr 16, 2019	IN PROGRESS
BDAA-73	7. Pilot	☑	Jan 23, 2019	Jan 23, 2019	OPEN
BDAA-74	8. Service Development & Deploy	☑	Jan 23, 2019	Feb 28, 2019	OPEN
	...tion	☑	Jan 23, 2019	Jan 23, 2019	OPEN
		☑	Jan 23, 2019	Jan 23, 2019	OPEN

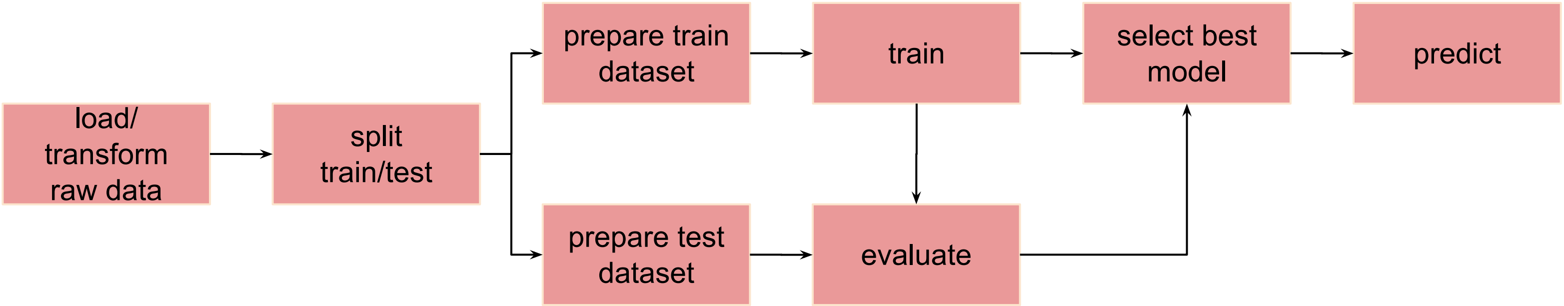
Project Involvement



JIRA
Bitbucket
Confluence
Atlassian Bamboo

Pipelines

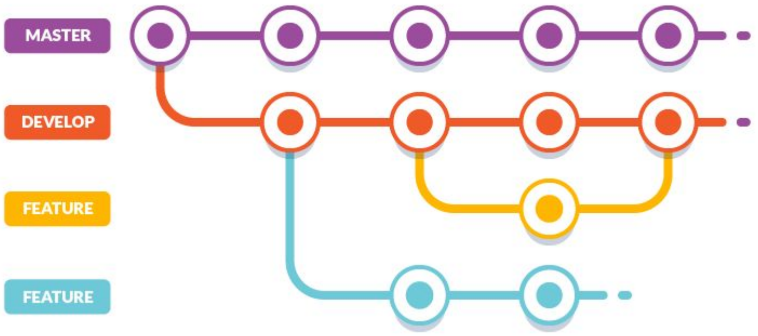
- Jupyter Notebooks for prototyping & visualization only
- End-to-end or selected steps
- YAML configs for all params
- One-button run



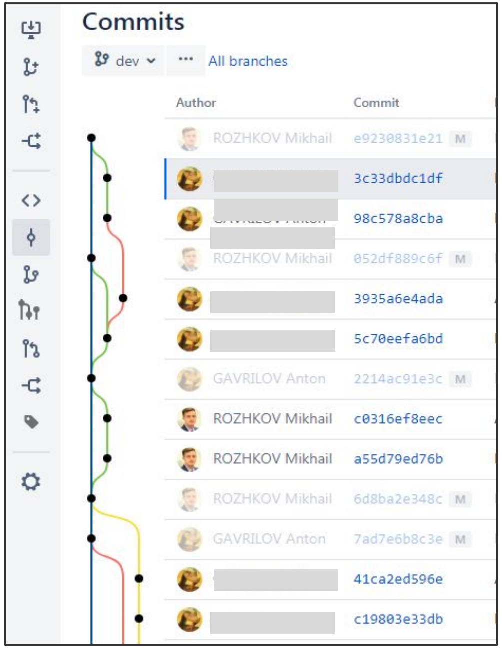
Code versioning and git-flow approach

ATLASSIAN Bitbucket

- Version control
- Re-usable .py modules
- Tests...

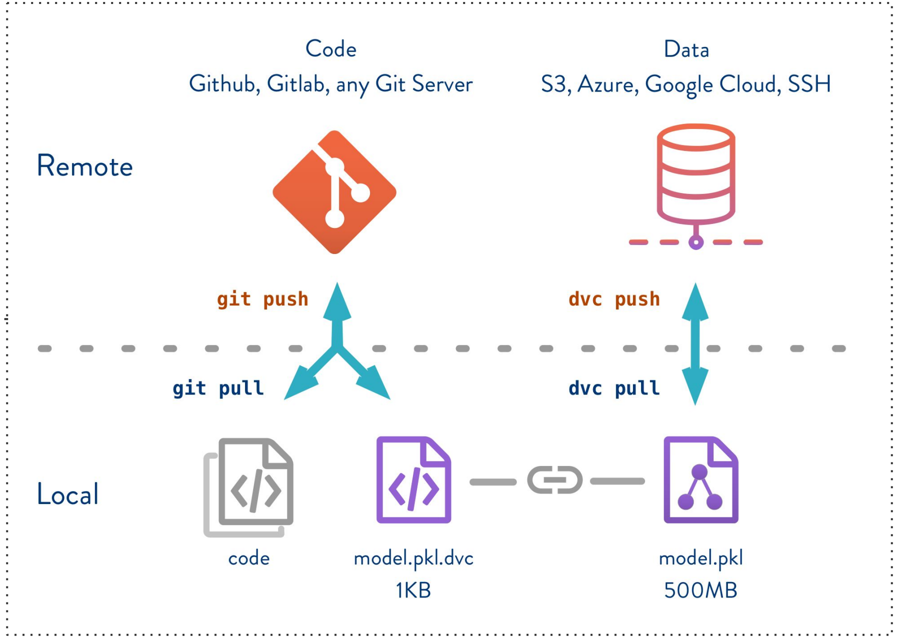


Source: https://www.bitbull.it/en/blog/how-git-flow-works/



Data and artifacts

- Version Control
- Shared remote storage
- Access



Experiments Management

- browse history
- compare results
- share results
- methodology and procedures

params

metrics

runs

	Date	User	Source	Version	Parameters: alpha	Parameters: l1_ratio	Metrics: mae	Metrics: r2	Metrics: rmse
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	1	0.649	0.04	0.862
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.5	0.648	0.046	0.859
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.2	0.628	0.125	0.823
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	1	0	0.619	0.176	0.799
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	1	0.648	0.046	0.859
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.5	0.628	0.127	0.822
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.2	0.621	0.171	0.801
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0	0.615	0.199	0.787
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	1	0.578	0.288	0.742
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.5	0.578	0.288	0.742
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.2	0.578	0.288	0.742
☐	2018-06-04 23:00:08	mlflow	train.py	05e956	0	0	0.578	0.288	0.742

Task: make it in right way

1. Try / select appropriate tools
2. Apply in your own project
3. Convince other DS to try / use
4. Share knowledge & help
5. Estimate economic value of changes
6. Plan / implement changes

Meetups with code demonstration and join projects work

Channels

- Internal meetups
- External meetups & conferences
- Real project code, statuses dashboard and documentation

Insights

- easy to convince people in your own team
- cross-teams collaboration show benefits of new approach

More projects and people

1. Predict an user behavior on Landing Page ✓
2. Client LifeTime Value prediction ✓
3. Virtual Assistant for a Call Center ✓

Task: make it in right way

1. Try / select appropriate tools
2. Apply in your own project
3. Convince other DS to try / use
4. Share knowledge & help
5. Estimate economic value of changes
6. Plan / implement changes

DS Brainstorm to estimate AS-IS and TO-BE practices

- 19 Data Scientists from different departments
- Estimate ~40 common tasks in ML projects
  - how much efforts spent (AS-IS), in man-days
  - what are opportunities and barriers
  - how to improve
  - estimated benefits (TO-BE), in man-days
- Group of tasks to estimate
  - Analyze & Plan
  - Get and prepare data
  - Train Model & Evaluate model
  - Productionize

Insights

- 90 % of problems are common for all teams
- Experimentation and Deployment stages have high potential for improvements
- ideas cover proposals for new tools, processes change, teams collaboration, education improvements

Tasks costs estimates are similar to Gartner's report

Task (Proportion of Effort)	Subtasks	Business	Stakeholder: Data Scientist	Stakeholder: IT/ Operations	Total (Gartner)
1. Problem Understanding (5% to 10%)	a) Determine Objective	X	X		5% to 10 %
	b) Define Success Criteria	X	X		
	c) Assess Constraints	X	X	X	
2. Data Understanding (10% to 25%)	a) Assess Data Situation	X	X	X	30% to 65 %
	b) Obtain Data (Access)		X	X	
	c) Explore Data	X	X	X	
3. Data Preparation (20% to 40%)	a) Filter Data		X	X	
	b) Clean Data		X	X	
	c) Feature Engineering	X	X		
4. Modeling (20% to 30%)	a) Select Model Approach		X		25% to 40 %
	b) Build Models		X		
5. Evaluation of Results (5% to 10%)	a) Select Model		X		
	b) Validate Model		X		
	c) Explain Model	X	X		
6. Deployment (5% to 15%)	a) Deploy Model		X	X	5% to 15 %
	b) Monitor and Maintain	X	X	X	
	c) Terminate	X	X	X	

Source: Gartner (January 2017)

https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/preparing_and_architecting_for_machine_learning.pdf

Estimated Value Potential

~ 40 - 60 % FTE cost reduction

~ 2 times faster projects

~ up to 50 % of efforts can be allocated to Modeling & Evaluation

Work in progress to upgrade tools for ML projects

- Design ML Platform (set of tools)
  - Feature Store
  - Model Registry
  - Experiments Management
  - Metadata Management
  - Deployment Management
  - etc.
- Cost-Benefits Analysis

Conclusions

1. ML projects need different approach and tools
2. Data and experiments are crucial
3. Fast experimenting and reproducibility are important
4. Estimated benefits from ML automation are convincing enough