

Машинное обучение: работа в команде и воспроизводимые эксперименты

Эмели Драль

Проектная работа

Весь объем работы можно разделить на **три** стадии:

- Предпроектное исследование
- **Работа над проектом**
- Работа после окончания проекта

Работа в команде и воспроизводимые эксперименты

1. Teamwork
2. Data & Code
3. Reproduceable Research

Teamwork

Teamwork

Teamwork

- Observability
- Communication
- Planning

Teamwork

Notion



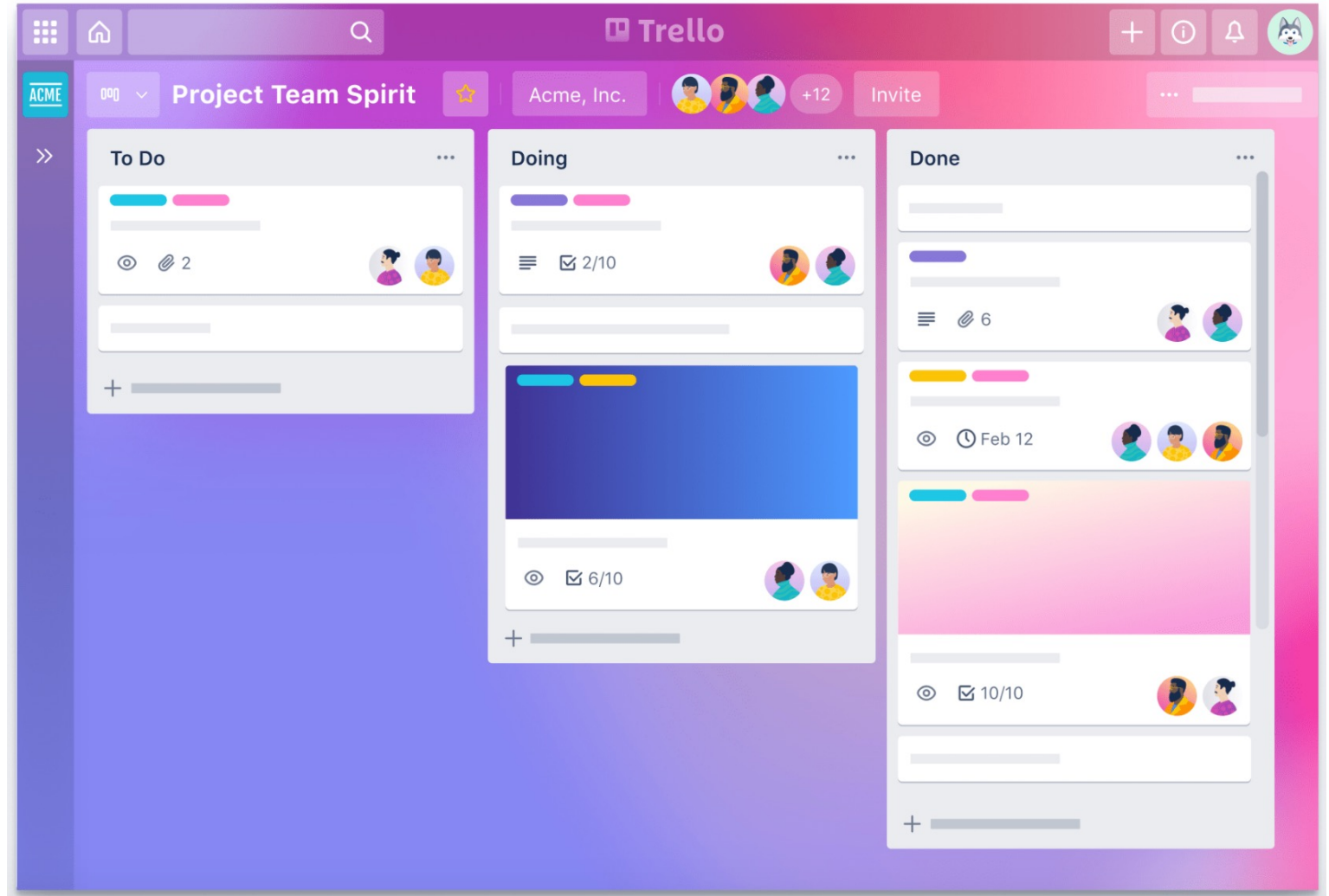
All-in-one workspace

One tool for your whole team. Write, plan, and get organized.

<https://www.notion.so/>

Trello

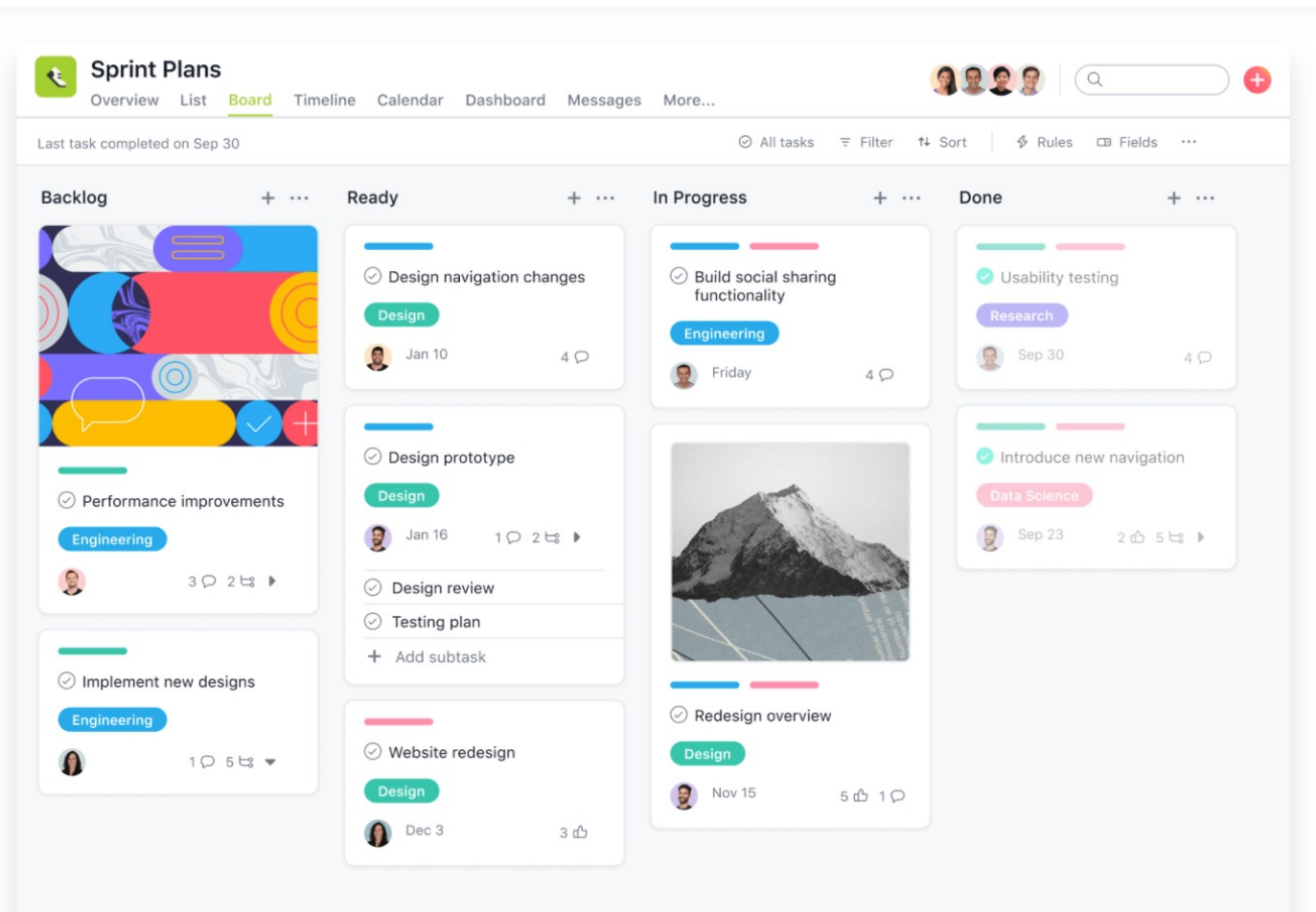
Teamwork



<https://trello.com/en>

Teamwork


Asana











[Why Asana? ▾](#)[Solutions ▾](#)[Resources ▾](#)[Enterprise](#)[Pricing](#)

<https://asana.com/uses/project-management>




Teamwork

Miro

miro | Brainwriting ☆ | 

Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6
Idea 1	Idea 2	Idea 3			
	Idea 2 improvement	Idea 3 improvement			

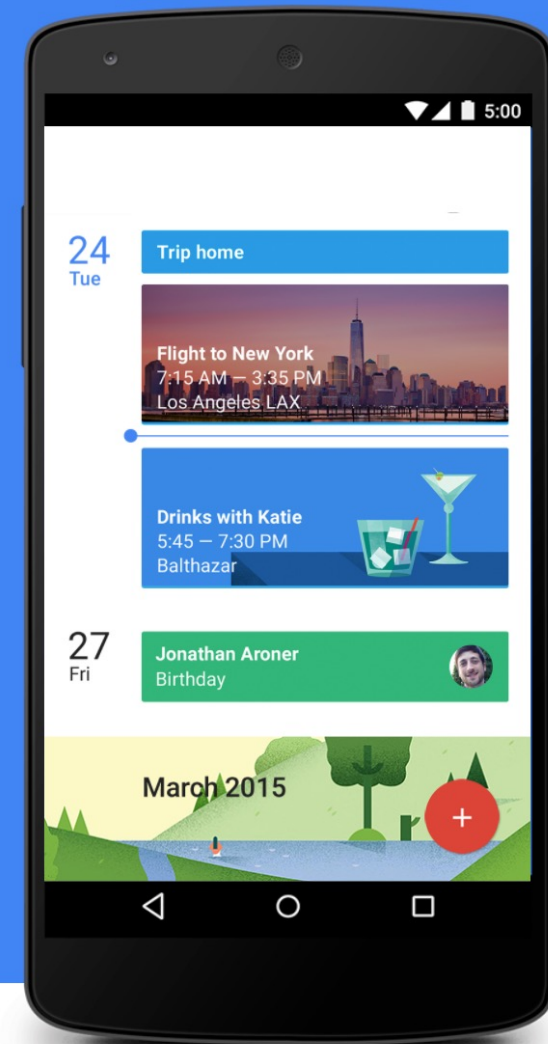
<https://miro.com/>

Teamwork

Calendar

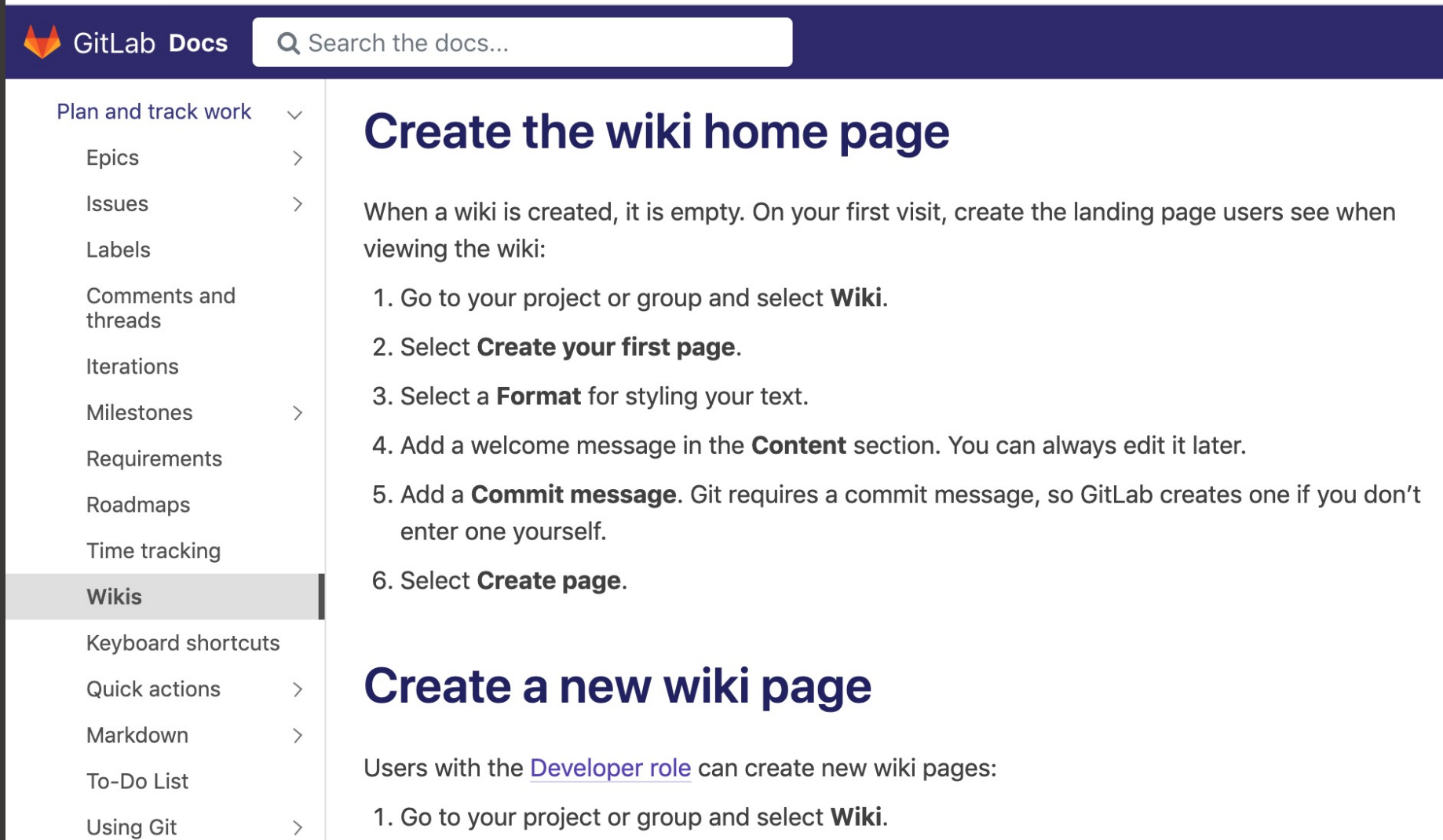
Make the
most of
every day

The new Google Calendar app helps
you spend less time managing your
schedule and more time enjoying it.



Teamwork

Gitlab/Github (engineers)



The screenshot shows the GitLab Docs interface. The sidebar on the left lists various documentation topics, with 'Wikis' highlighted. The main content area displays the article 'Create the wiki home page', which explains that a new wiki is empty and provides a six-step guide to setting up the home page. The steps include selecting the 'Wiki' option, creating a first page, choosing a format, adding a welcome message, adding a commit message, and finally creating the page.

Plan and track work ▾

- Epics >
- Issues >
- Labels
- Comments and threads
- Iterations
- Milestones >
- Requirements
- Roadmaps
- Time tracking
- Wikis**
- Keyboard shortcuts
- Quick actions >
- Markdown >
- To-Do List
- Using Git >

Create the wiki home page

When a wiki is created, it is empty. On your first visit, create the landing page users see when viewing the wiki:

1. Go to your project or group and select **Wiki**.
2. Select **Create your first page**.
3. Select a **Format** for styling your text.
4. Add a welcome message in the **Content** section. You can always edit it later.
5. Add a **Commit message**. Git requires a commit message, so GitLab creates one if you don't enter one yourself.
6. Select **Create page**.

Create a new wiki page

Users with the [Developer role](#) can create new wiki pages:

1. Go to your project or group and select **Wiki**.

<https://docs.gitlab.com/ee/user/project/wiki/>

Data & Code

Data

Data and Datasets

- Store data with the metainformation
- Use Data Storages, Data Bases, Drives
- Be careful with storing data in the repositories
- Use Data Version Control for datasets (if applicable)

Data

Data and Datasets

- Store data with the metainformation
- Use Data Storages, Data Bases, Drives
- Be careful with storing data in the repositories
- Use Data Version Control for datasets (if applicable)

Anti-patterns:

- Locally stored files
- Attaches in mail

Data

Data and Datasets



FEATURES

DOC

Open-source
Version Control System
for Machine Learning Projects



Download
(Mac OS)





Watch video
How it works

Analytical code




Code


← → ↻ github.com/evidentlyai/evidently









 Search or jump to... / Pull requests Issues Marketplace Explore

 [evidentlyai](#) / [evidently](#)

<> Code Issues 15 Pull requests 3 Actions Projects Wiki Security Insights Settings

 main ▾  8 branches  5 tags Go to file Add file ▾ ↓ Code ▾

 **emeli-dral** Updated statistical test for categorical features 7893fc4 23 hours ago ⌚ 165 commits

 evidently	Updated statistical test for categorical features	23 hours ago
 .gitignore	Add __pycache__ to gitignore	last month
 LICENSE	Updated license	8 months ago
 README.md	Update README.md	6 days ago
 config.json	Renamed production to current.	last month
 config.yaml	Renamed production to current.	last month
 setup.py	Add pyyaml to required.	last month
 setupbase.py	Initial version	8 months ago

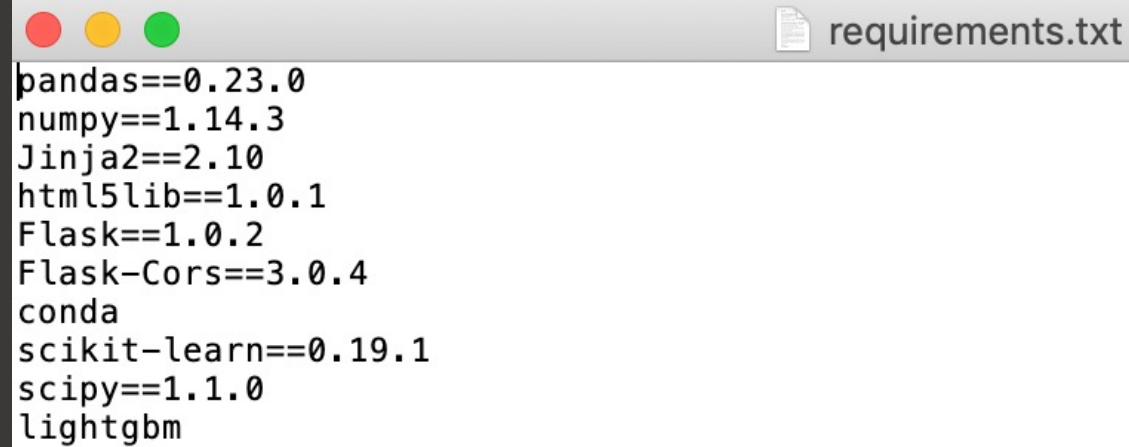
Code

Analytical code

- Use control version systems
- Save imperfect code (even if it does not work properly)
- Do not transfer files by messengers, email
- Do not store your code in drives, folders

Environment & Requirements

Code



```
pandas==0.23.0
numpy==1.14.3
Jinja2==2.10
html5lib==1.0.1
Flask==1.0.2
Flask-Cors==3.0.4
conda
scikit-learn==0.19.1
scipy==1.1.0
lightgbm
```

Code

Environment & Requirements

- Use control version systems
- Save imperfect code (even if it does not work properly)
- Do not transfer files by messengers, email
- Do not store your code in drives, folders

Code

Environment & Requirements

- Create `requirements.txt`
- For python use `virtualenv`:
 - to create: `python3 -m venv venv`
 - to activate: `. venv/bin/activate`
 - to deactivate: `deactivate`
- Also consider services like Docker, Kubernetes, MLFlow

Reproducible Experiments

Reproducible
Experiments

Experiments & results

- Document data sources and data preprocessing process

Reproducible Experiments

Experiments & results

- Document data sources and data preprocessing process
- For complex ETL operations use Pipelines, workflow managers

Reproducible Experiments

Experiments & results

- Apache Airflow
- Luigi
- Prefect
- Argo
- KubeFlow
- MLFlow

Reproducible
Experiments

Experiments & results

Apache Airflow

Airflow is a platform created by the community to programmatically author, schedule and monitor workflows.

Install

<https://airflow.apache.org/>

Reproducible Experiments

Experiments & results

- Document data sources and data preprocessing process
- For complex ETL operations use Pipelines, workflow managers
- Write docstrings and comments for functions, classes, modules

Reproducible Experiments

Experiments & results

- Document data sources and data preprocessing process
- For complex ETL operations use Pipelines, workflow managers
- Write docstrings and comments for functions, classes, modules
- Make sure using deterministic algorithms

Reproducible Experiments

Experiments & results

- Document data sources and data preprocessing process
- For complex ETL operations use Pipelines, workflow managers
- Write docstrings and comments for functions, classes, modules
- Make sure using deterministic algorithms
- For randomized algorithms use RANDOM STATE (SEED)

Reproducible Experiments

Experiments & results

- Document data sources and data preprocessing process
- For complex ETL operations use Pipelines, workflow managers
- Write docstrings and comments for functions, classes, modules
- Make sure using deterministic algorithms
- For randomized algorithms use RANDOM STATE (SEED)
- Log experiment together with artifacts: metrics, datasets, models

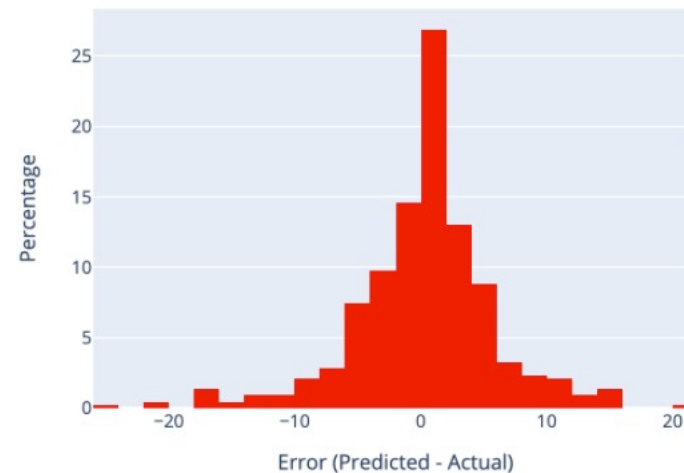
Reproducible Experiments

Experiments & results

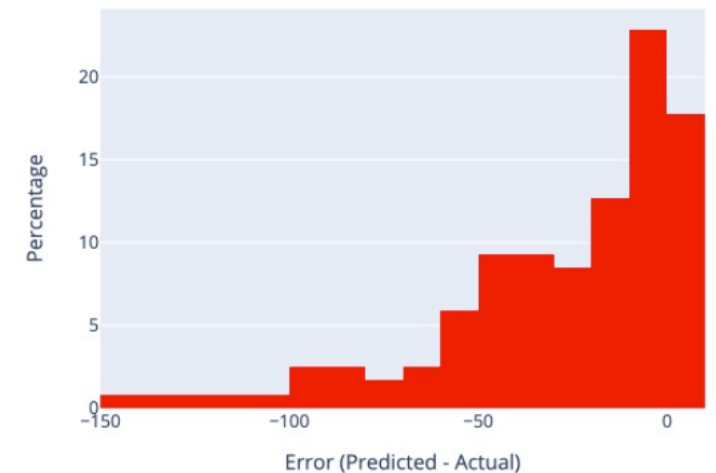
4. Regression Model Performance

Analyzes the performance of a regression model and model errors.

Reference: Error Distribution



Current: Error Distribution



<https://github.com/evidentlyai/evidently>

https://github.com/evidentlyai/evidently/blob/main/evidently/tutorials/ibm_hr_attrition_model_validation.ipynb

Reproducible Experiments

Experiments & results

Data Drift Evaluation with Evidently

 Track machine learning training runs in an experiment. [Learn more](#)

Experiment ID: 3

Artifact Location: file:///Users/emeli/Dev/evidently/mlflow/examples/evidently/mlruns/3

▼ Notes [🔗](#)

None

Search Runs:

Filter

Search

Clear





Showing 6 matching runs

Compare

Delete

Download CSV 

Columns

<input type="checkbox"/>	Start Time	Run Name	User	Source	Version	Models	Parameters		Metrics >		
							begin	end	atemp	holiday	humidity
<input type="checkbox"/>	🟢 2021-07-13 20:05:35	-	emeli	 ipykernel_l	-	-	2011-02...	2011-02...	0	0	0
<input type="checkbox"/>	🟢 2021-07-13 20:05:35	-	emeli	 ipykernel_l	-	-	2011-02...	2011-02...	0	0	0
<input type="checkbox"/>	🟢 2021-07-13 20:05:35	-	emeli	 ipykernel_l	-	-	2011-01...	2011-02...	0	0	0
<input type="checkbox"/>	🟢 2021-07-13 20:05:35	-	emeli	 ipykernel_l	-	-	2011-01...	2011-01...	1	1	1

<https://www.mlflow.org/>

https://github.com/evidentlyai/evidently/blob/main/evidently/tutorials/mlflow_integration.ipynb

Работа в команде и воспроизводимые эксперименты

1. Teamwork
2. Data & Code
3. Reproduceable Research

Машинное обучение: работа в команде и воспроизводимые эксперименты

Спасибо!
Эмели Драль