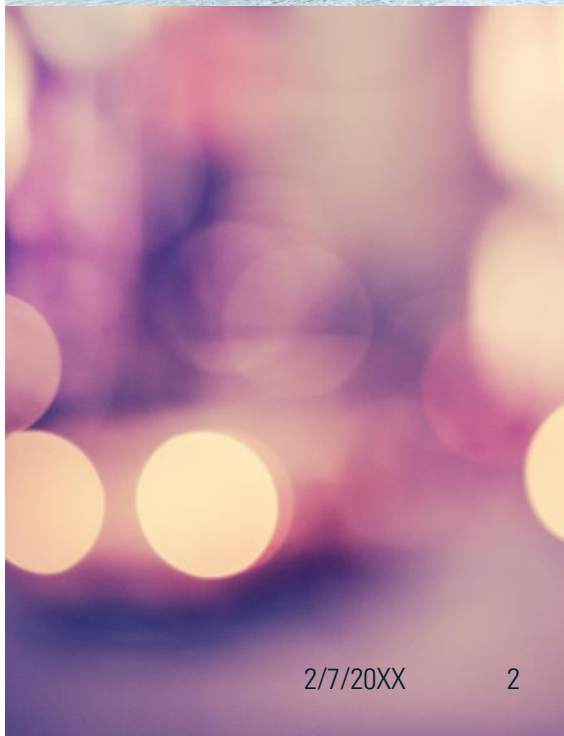
AMES HOUSING PRICE PREDICTION

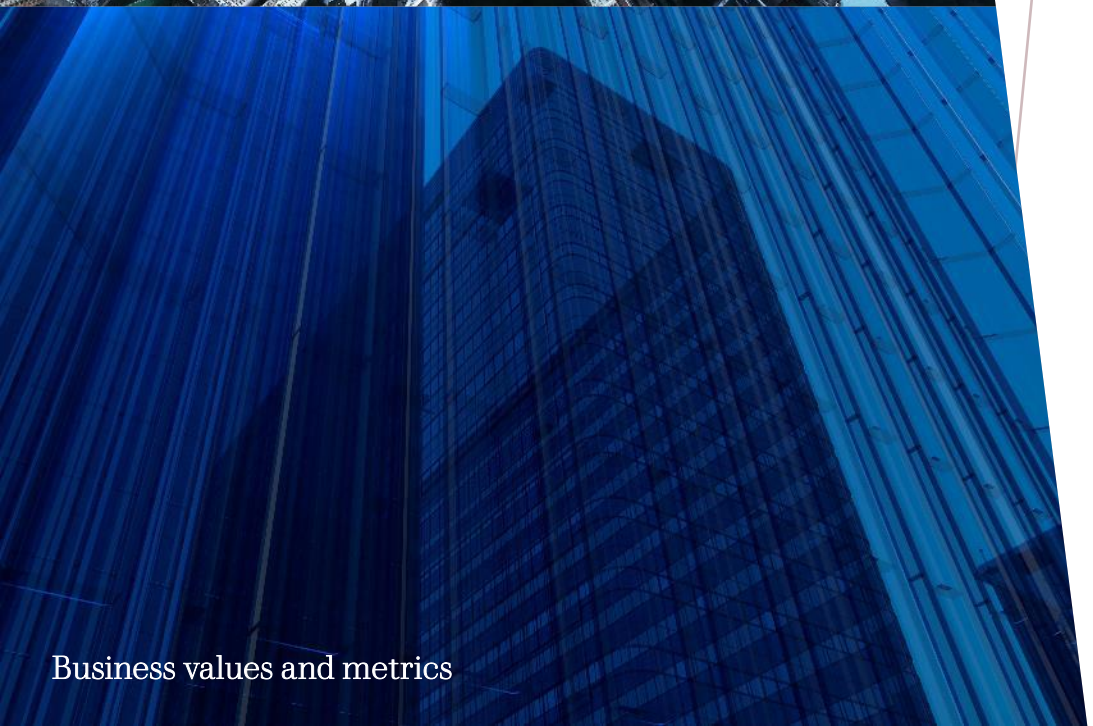
Dan Chen

2/20/2023

STRUCTURE

- Business values and metrics
- Modeling
- Pipeline
- API
- Q&A



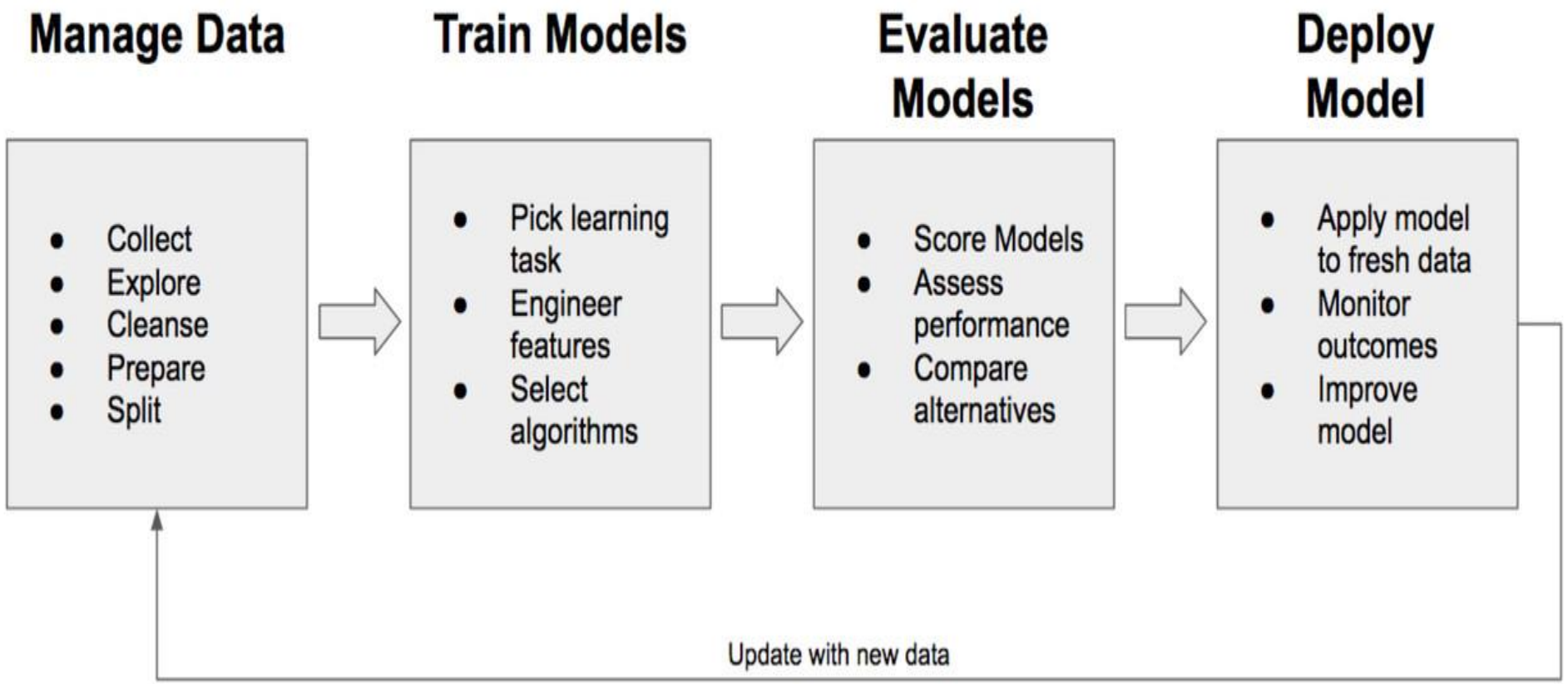


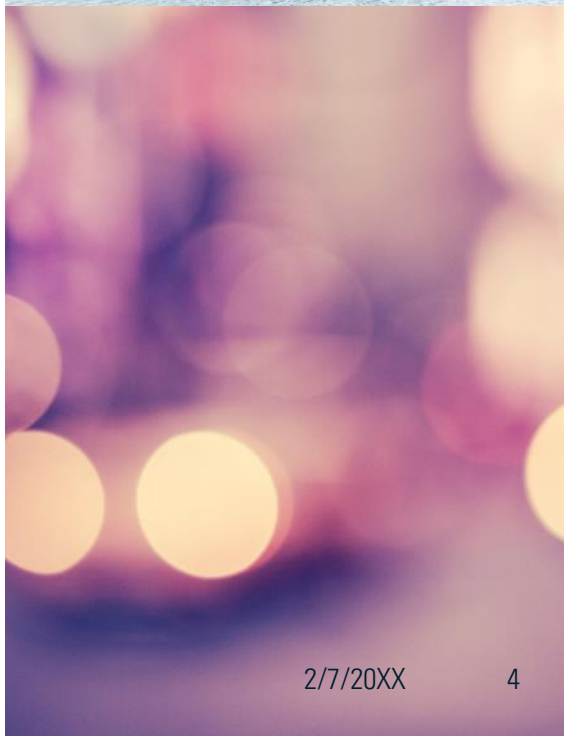
INTRODUCTION

Accurate housing price prediction help investors identify potential investment opportunities in real estate business

The aim of the project is to build a machine learning model pipeline to predict the sale price of homes

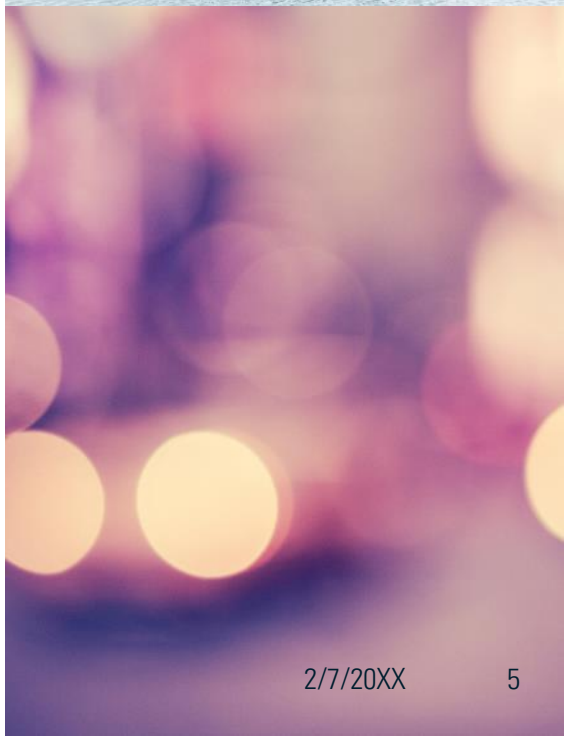
Machine Learning Modeling Cycle





MACHINE LEARNING PIPELINE

1. Data Analysis
2. Feature Engineering
3. Feature Selection
4. Model Training
5. Predictions Scoring
6. API
7. CI/CD
8. DOCKER

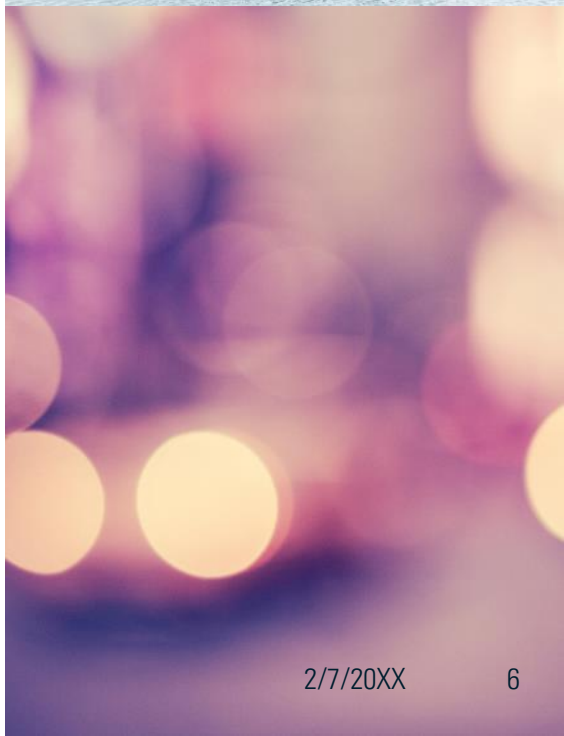


DATA ANALYSIS

Train.csv / Test.csv

The house price dataset contains 2051 rows, 80 columns:

- 79 predictive variables
- 1 target: SalePrice

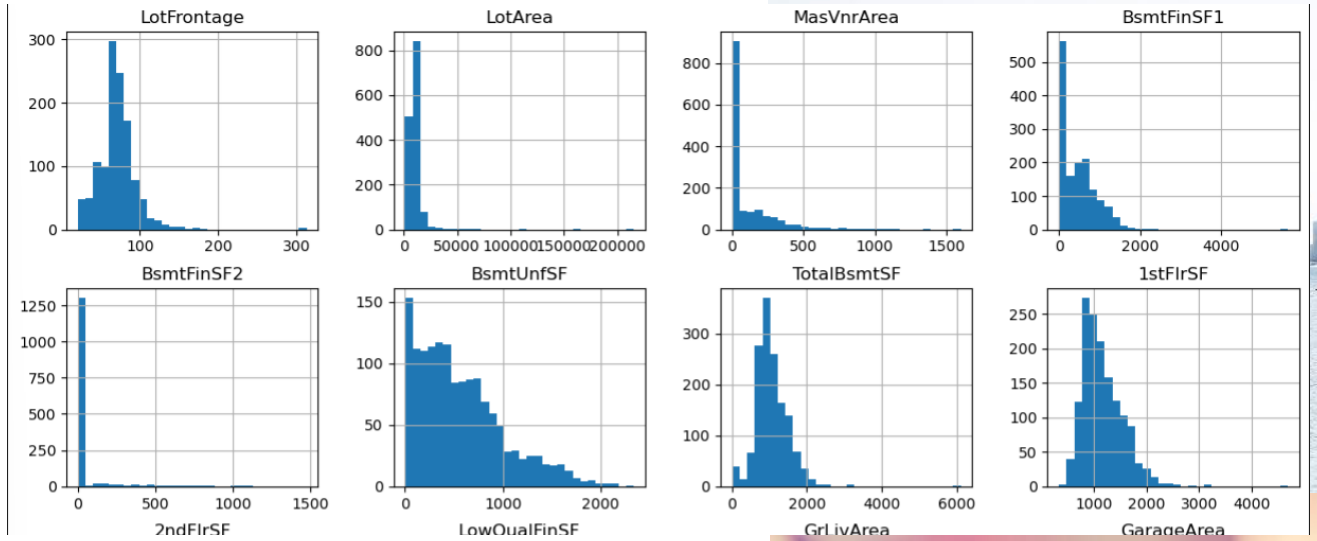


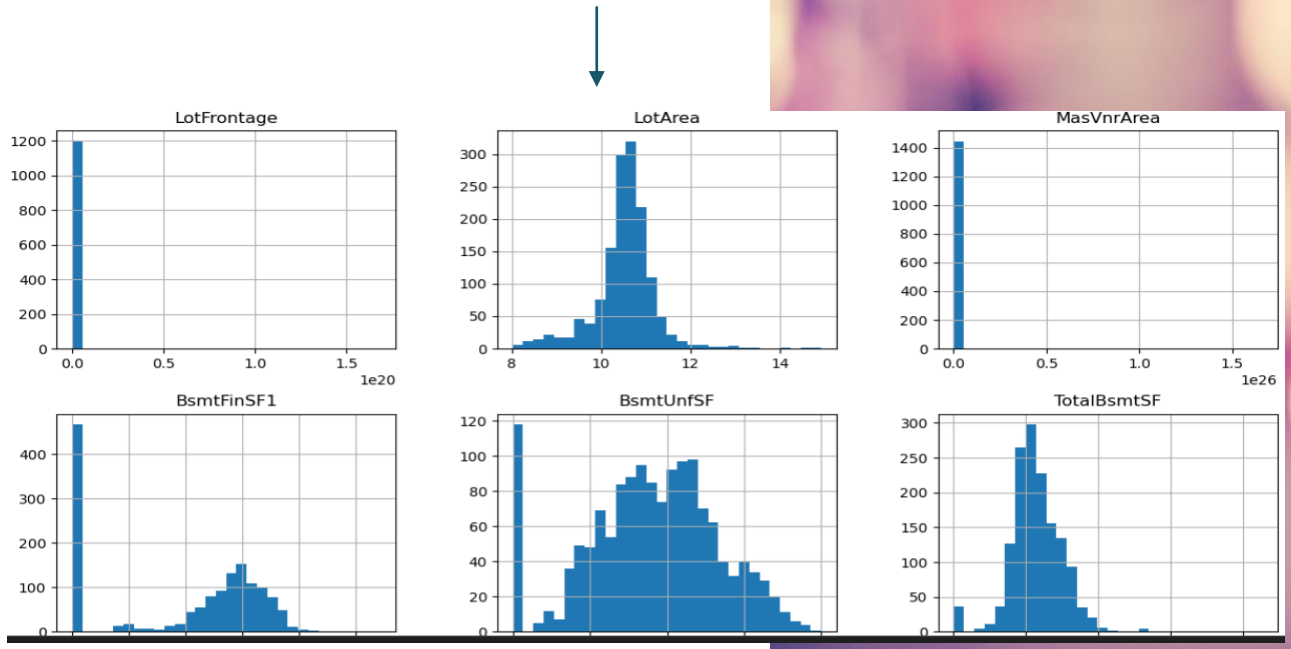
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview/description

DATA ANALYSIS

DATA TRANSFORMATION

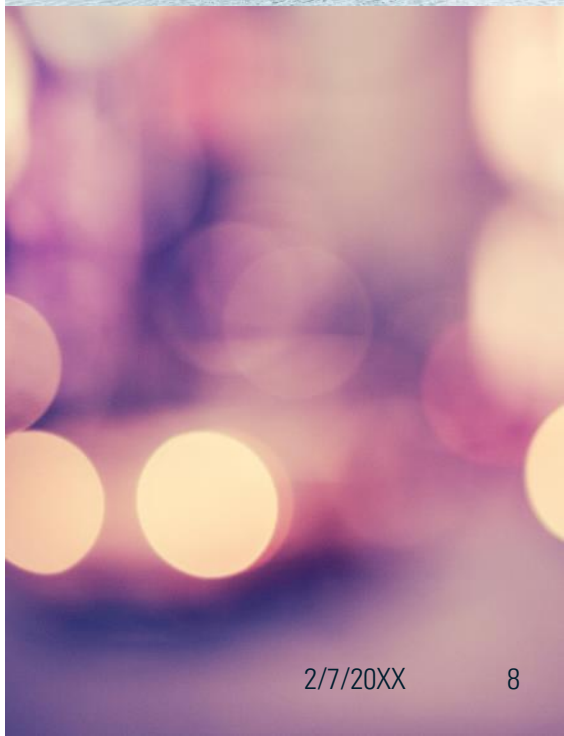
- Yeo-Johnson
- Log
- Skewed >> binary





FEATURE ENGINEERING

1. Missing values
2. Temporal variables
3. Non-Gaussian distributed variables
4. Categorical variables: remove rare labels
5. Categorical variables: convert strings to numbers
5. Put the variables in a similar scale
6. Mapping – assigned order
7. Feature scaling

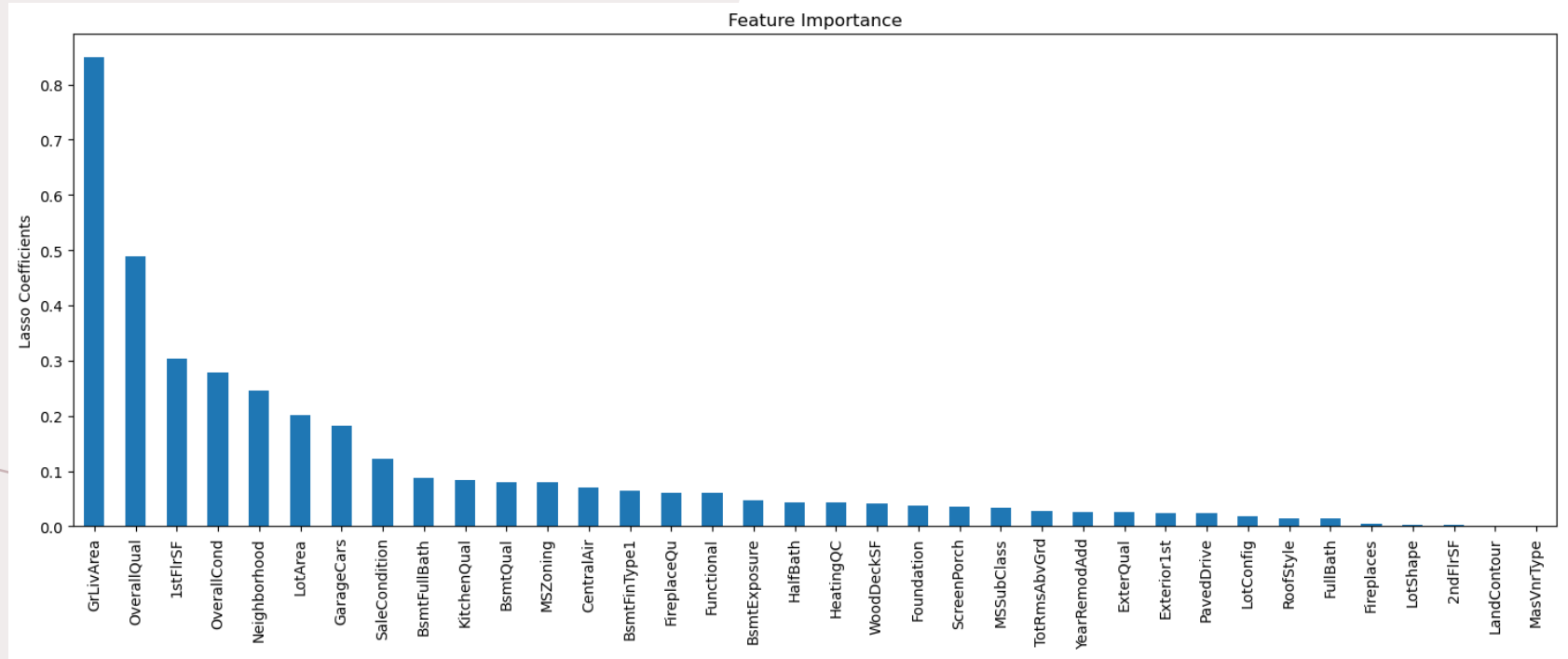


FEATURE SELECTION

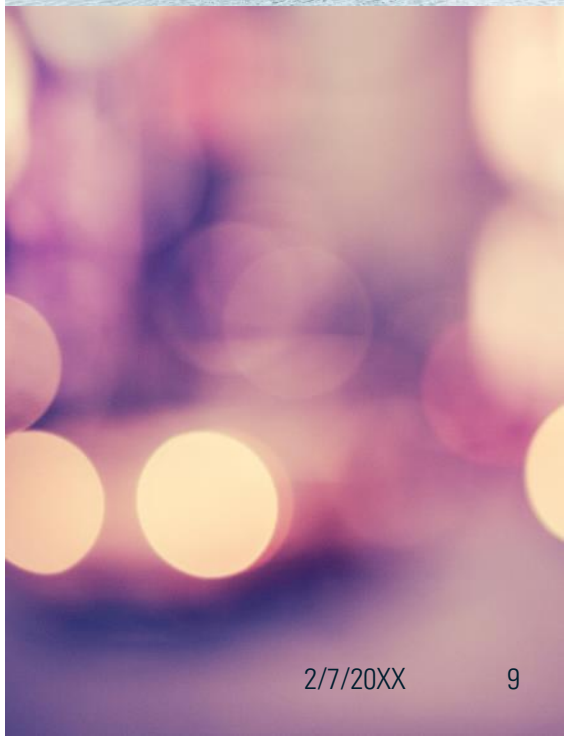
Use scikit learn: selectfrommodel

total features: 80

selected features: 36



https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

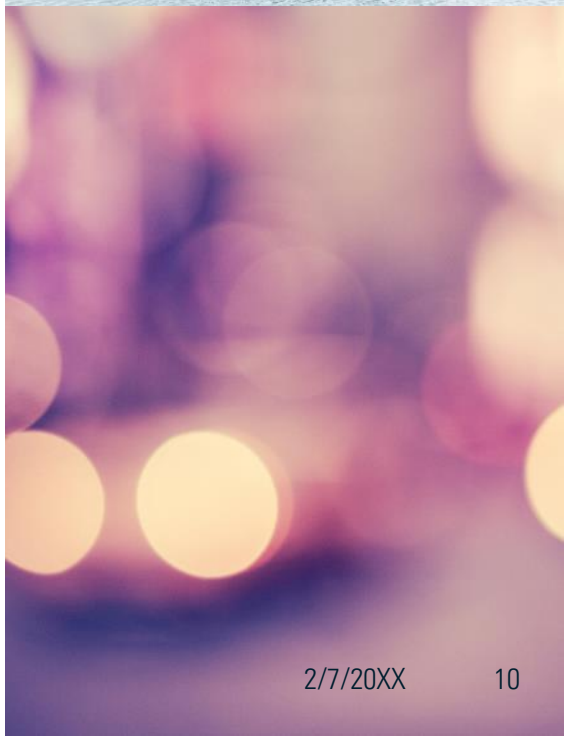


MODEL TRAIN

RMSE: 18179.7065

Model comparisons

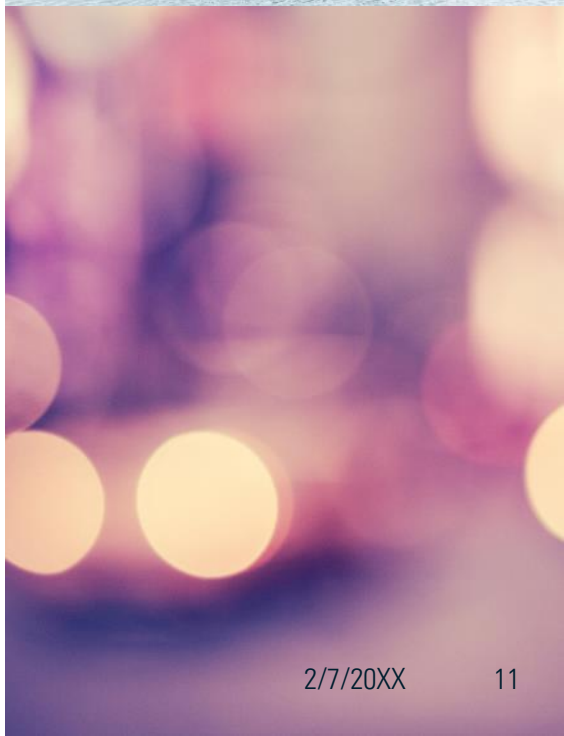
- Linear regression
- Lasso
- Ridge
- ElecticNet
- Random Forest
- XGBoost



https://github.com/chendbox/ames_housing_prediction

SCORING

	TRAIN	TEST
RMSE	27769	33016
R2	0.876496042007156	0.8413724432687755

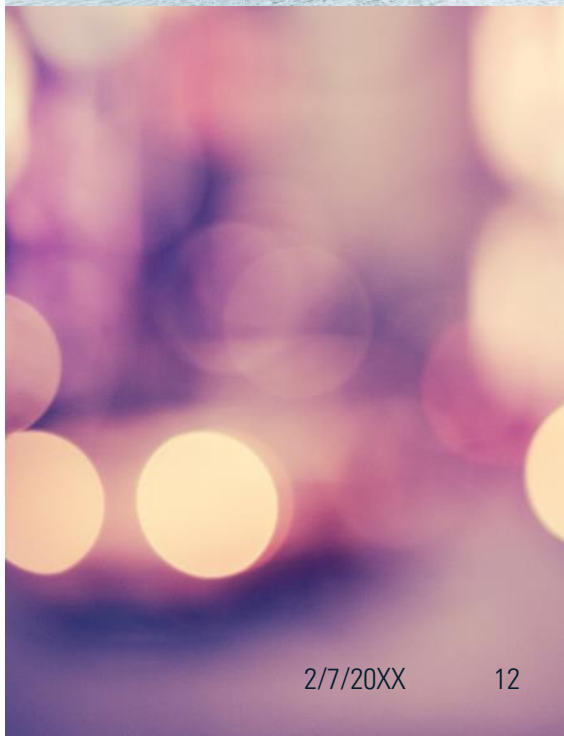


LINEAR REGRESSION

Model_pipeline.py

- Feature pipeline
- Linear_model
- Score test data

```
joblib.dump(lin_model,
  'linear_regression.joblib')
```



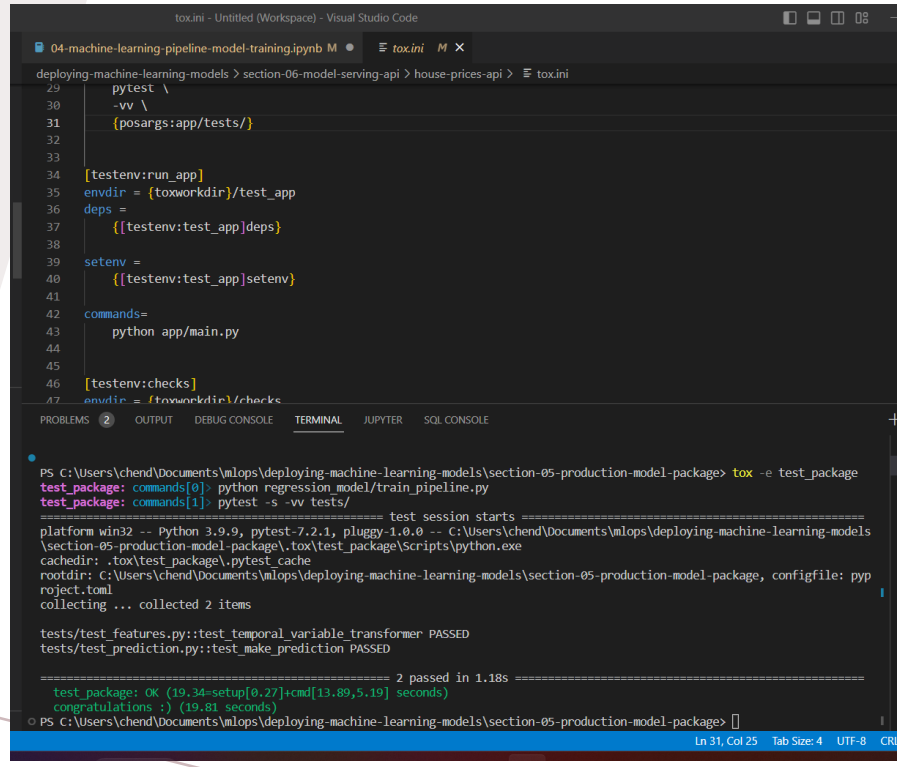
FAST API

Easy to implement

Product codes being tested with TOX https://tox.wiki/en/latest/

http://localhost:8001/

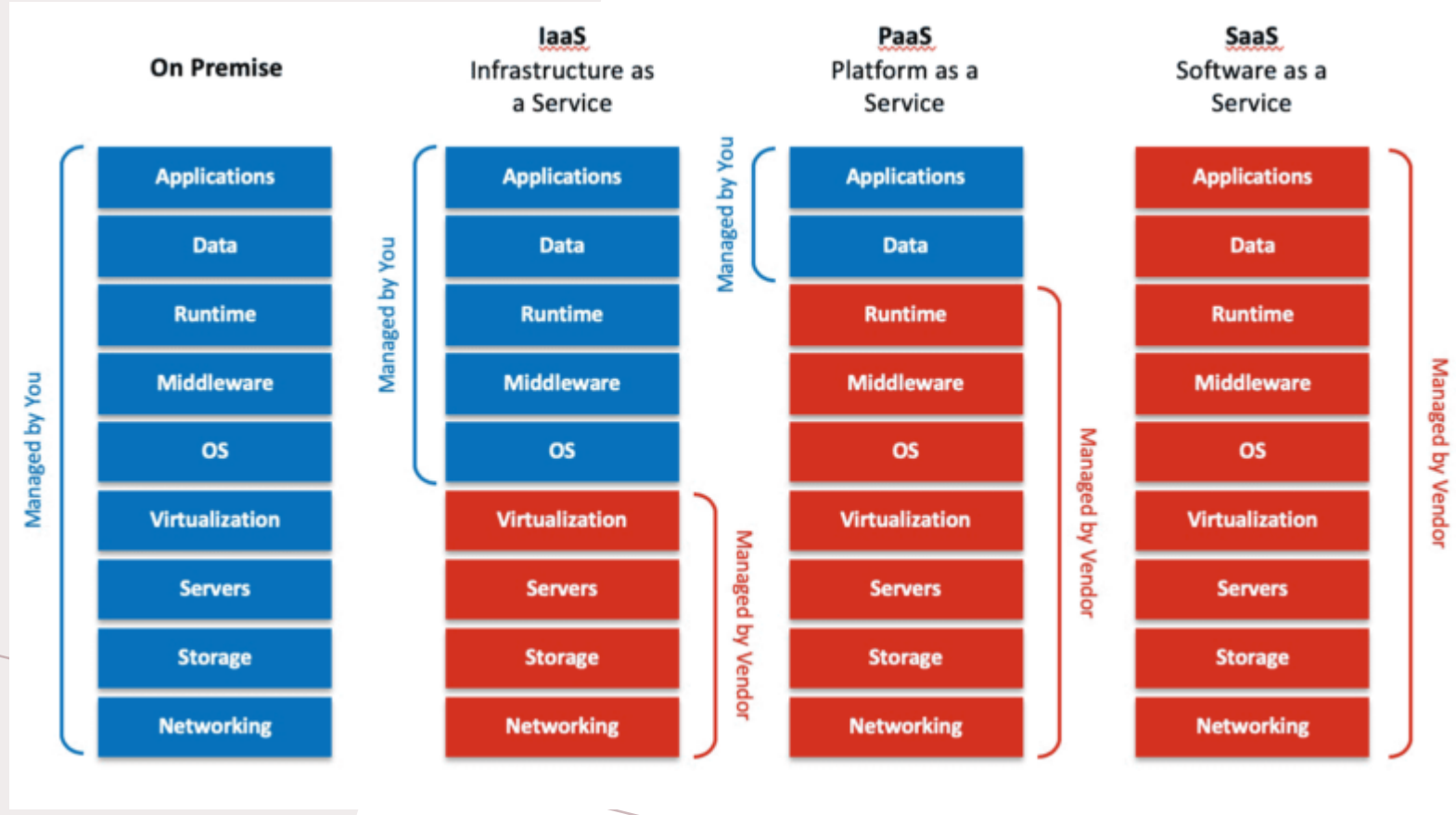
Demo



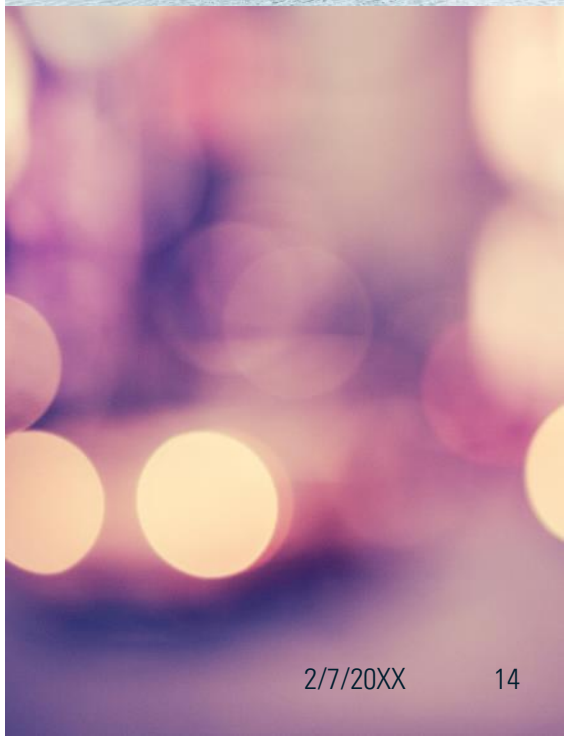
Local host

CLOUD ENVIRONMENT

Platform as a service (PaaS)



https://www.parallels.com/blogs/ras/iaas-vs-paas/



RAILWAY

Railway is an infrastructure platform where you can provision infrastructure, develop with that infrastructure locally, and then deploy to the cloud

- 500 hours free
- Support large dependency size on free tier
- Work with docker
- Github integration

Railway




railway.app/project/a69476c8-045a-4a81-9123-febefa88b85c/service/e65d2aad-9041-4a3cc-b31d-ba5511868038?id=dd0a00a6-d90e-4524-bf2d-7830ada0bc33&

/ gifted-comparison / production

Docs Help Starter Plan $ 4.98 487 Hrs

gifted-comparison
gifted-comparison-producti...

2 minutes ago via Up

dd0a00a Feb 20, 2023 12:00 pm

ACTIVE gifted-comparison-production.up.railway.app

Details Build Logs Deploy Logs

Filter logs using "", (), AND, OR, -

No deployment logs. All output sent to stdout will be displayed here

Showing 0 logs

Log Preferences

70°F
Partly sunny

Search

ENG

12:03 PM
2/20/2023

CI/CD

CircleCI is a continuous integration and continuous delivery platform that can be used to implement DevOps practices

Alternatives

- Jenkins
- Gitlab ci
- Travis CI




CI/CD

CircleCI is a continuous integration and continuous delivery platform that can be used to implement DevOps practices

Alternatives

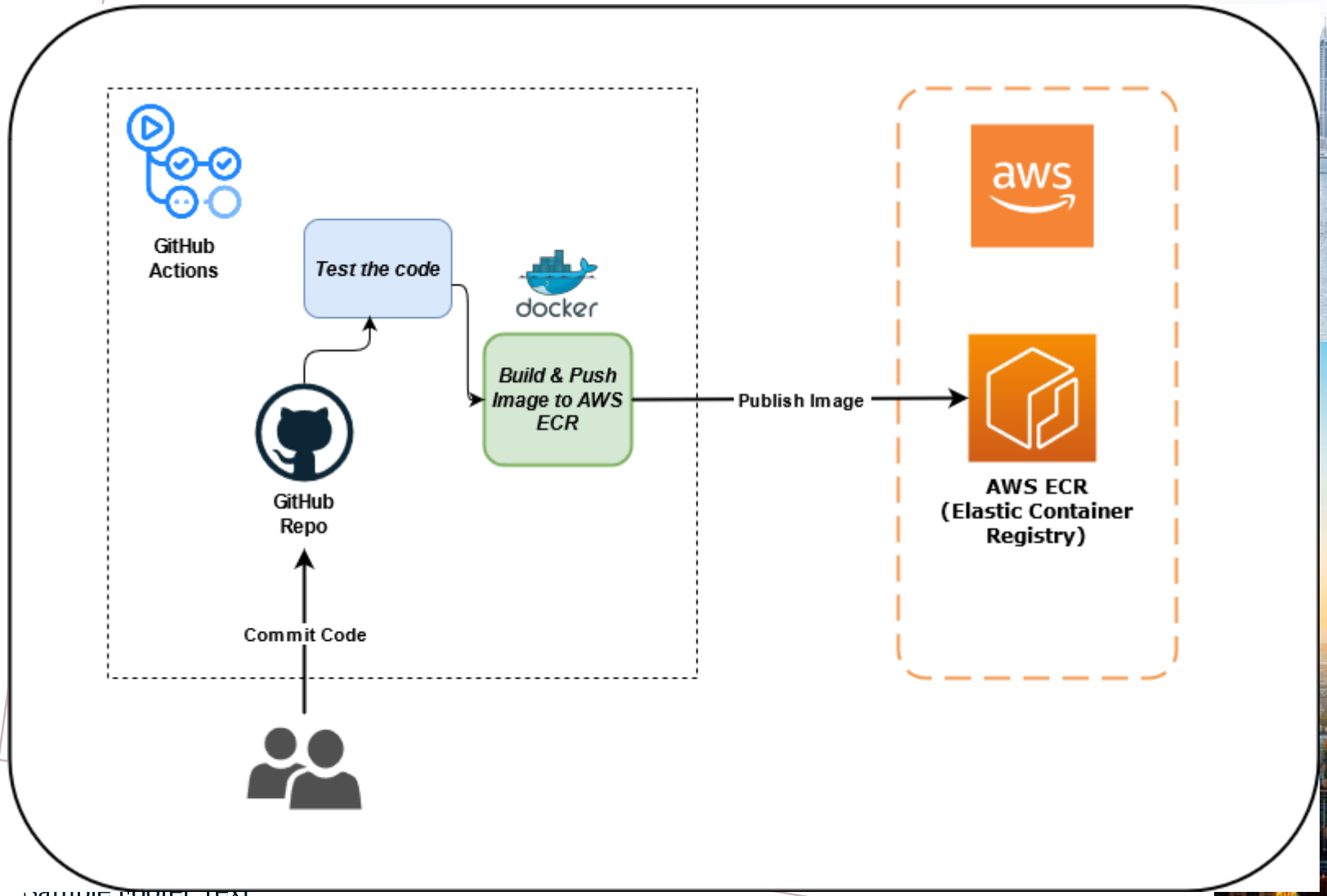
- Jenkins
- Gitlab ci
- Travis CI




IAAS

Infrastructure as a Service (IaaS) is a business model that delivers IT infrastructure like compute, storage, and network resources on a pay-as-you-go basis over the internet

AWS ENVIRONMENT



OUTCOME

1. Built a prediction model with low RMSE that can predict price with new data
2. Deploy the model serve from FAST API
3. Deploy model in production environment – RAILWAY
4. CI/CD to cloud platform with Docker




FAQ