Linear Regression

01418362 Introduction to Machine Learning

by 6610402230 Sirisuk Tharntham

Agenda

- Our Dataset
- Preprocessing
- Linear Regression
- Regularization
- Ensemble
- Kernelization

Our Dataset

https://www.kaggle.com/datasets/denkuznetz/taxi-price-prediction

Taxi Price Regression 🚕

Predict Taxi Price with Realistic Data and Hidden Patterns

k kaggle.com

shape 1000 x 11

- 1000 columns
- 11 rows (1 target)

Key Features

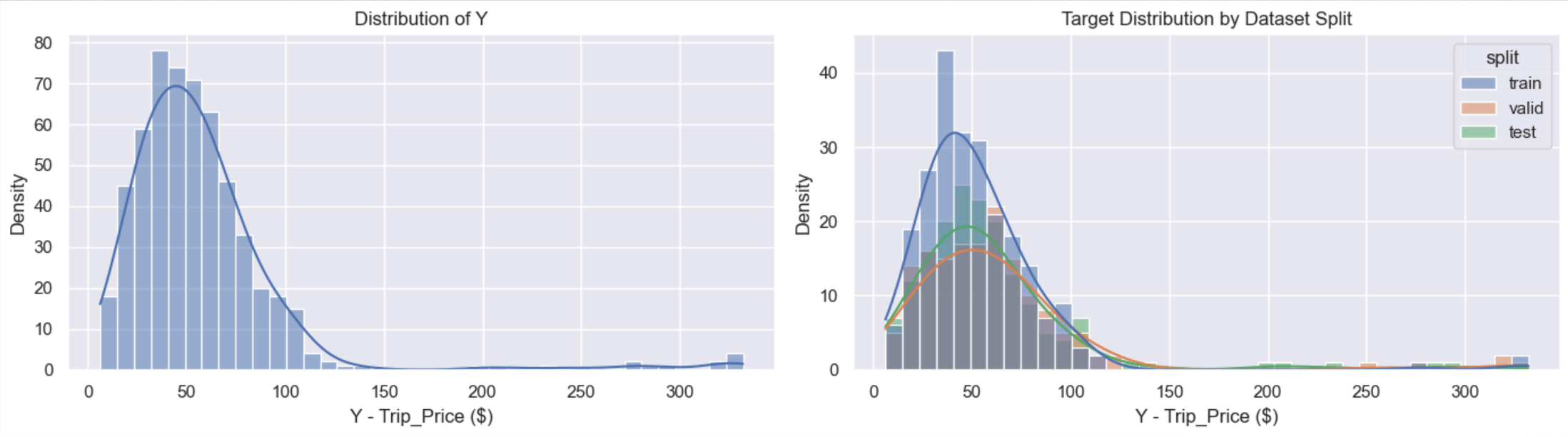
- Distance (in kilometers): The length of the trip.
- Pickup Time: The starting time of the trip.
- Dropoff Time: The ending time of the trip.
- Traffic Condition: Categorical indicator of traffic (light, medium, heavy).
- Passenger Count: Number of passengers for the trip.
- Weather Condition: Categorical data for weather (clear, rain, snow).
- Trip Duration (in minutes): Total trip time.
- Fare Amount (target): The cost of the trip (in USD).

Preprocessing

1. Drop null
2. Categorical normalization
    - to number
3. Standard scaler
4. Split train validation and test
    - Test 30%
    - Training 42%
    - Validation 28%

Final Dataset X,Y : 562 x 10 , 562 x 1

Preprocessing cont.



The distribution graph of trip price show value is about 50 USD

Linear Regression

- Implement classic linear regression model
- Training and tuning parameter
- Evaluation

Linear Regression cont.

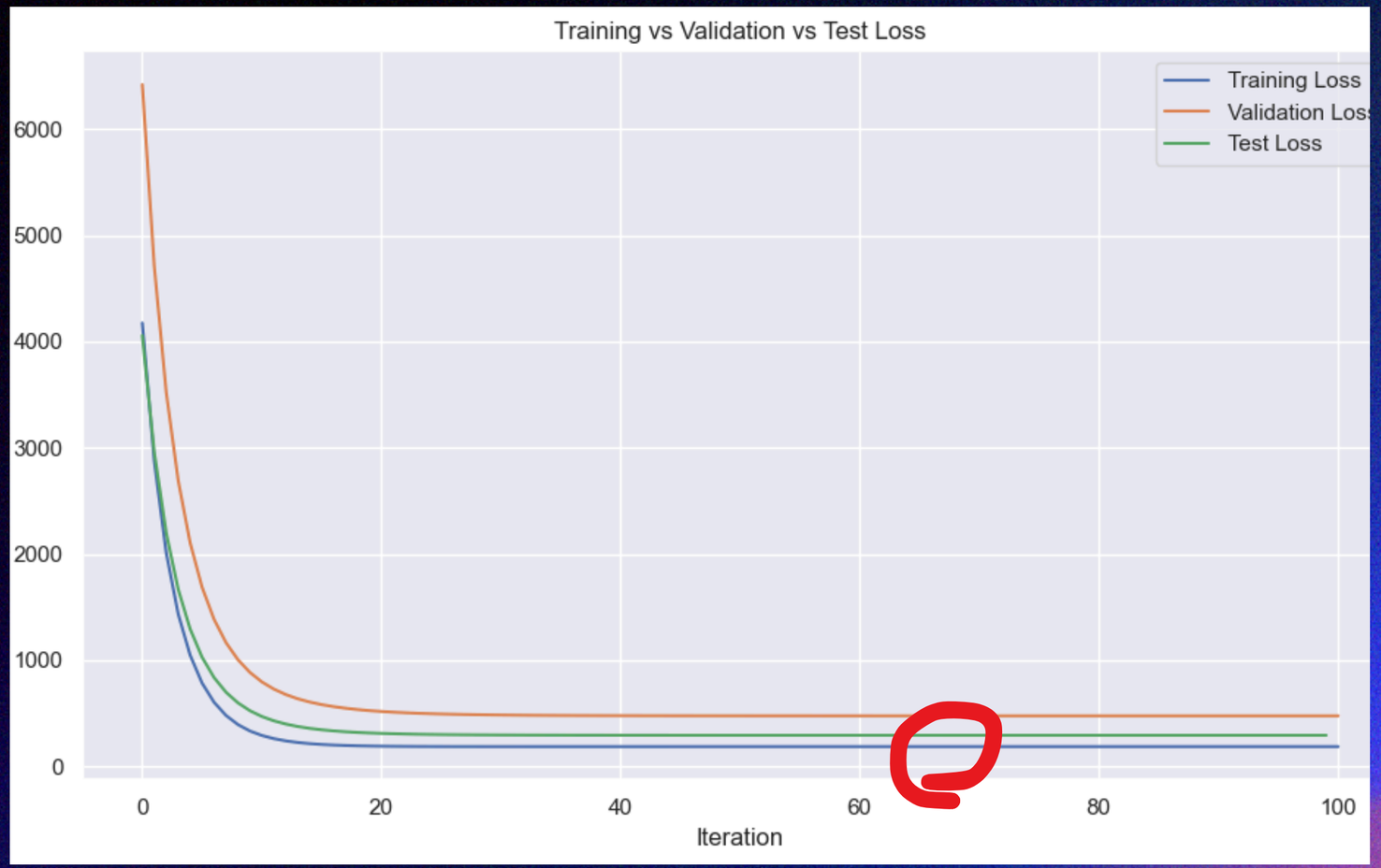
MSE Loss is 231.33
MAE Loss is 9.08

training loss below testing loss show that there may be varience problem

Regularization

- As the variance problem I try to solve with regularization
- Implement linear regression model with L1 Regularizer (Lasso)
- Training and tuning parameter
- Evaluation

Regularization cont.



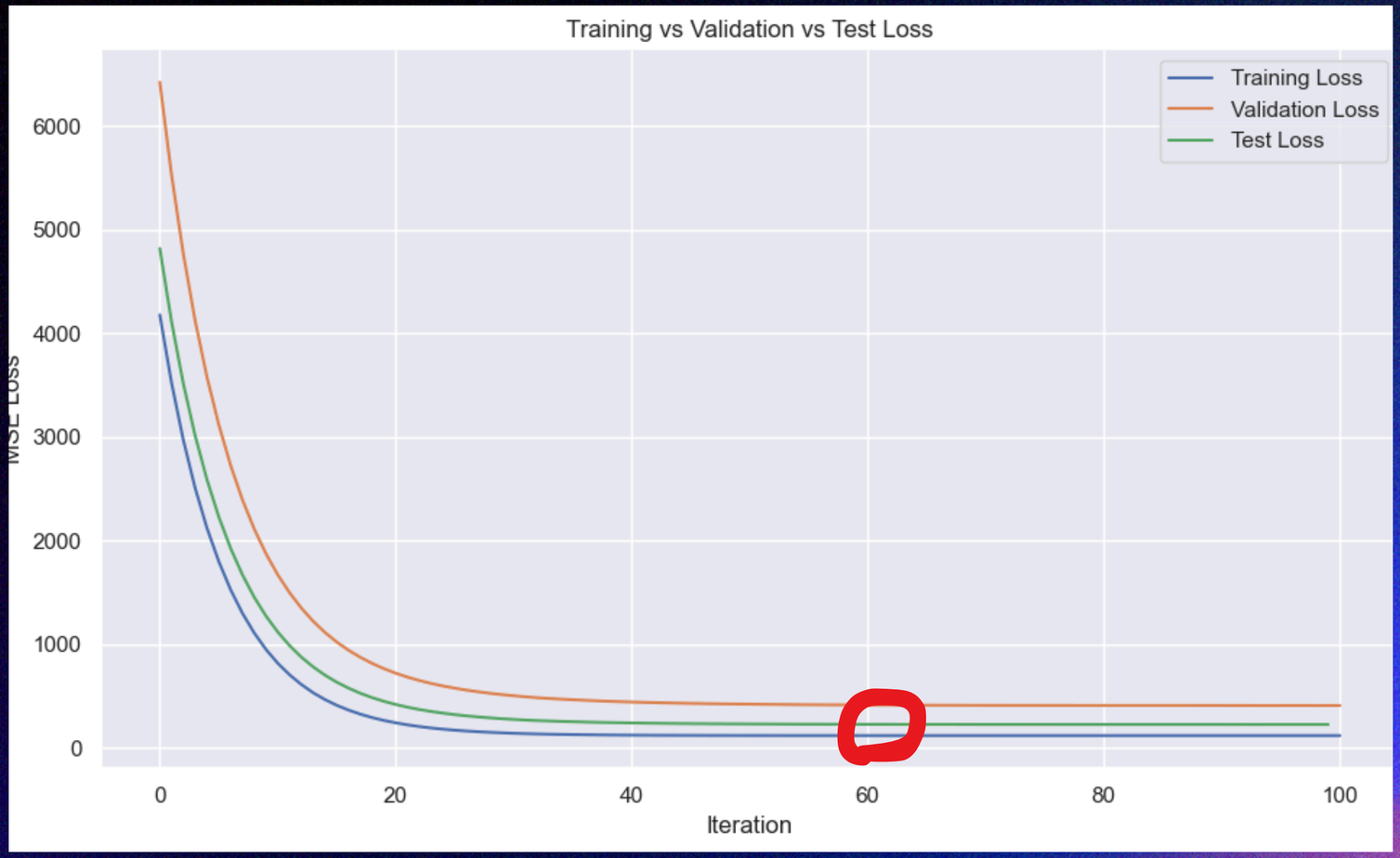
MSE Loss is 294.98
MAE Loss is 76.54

training loss below testing loss - that's still the same as previous indicated may be the bias problem

Ensemble

- From the previous I’ve inspect the loss and decide that bias problem
- So Next I will the gradient boosting (AnyBoost) with Lasso
- Implement same Laso algorithm
- Add implementation of the gradient boosting
- Training and tuning parameter
- Evaluation

Ensemble cont.

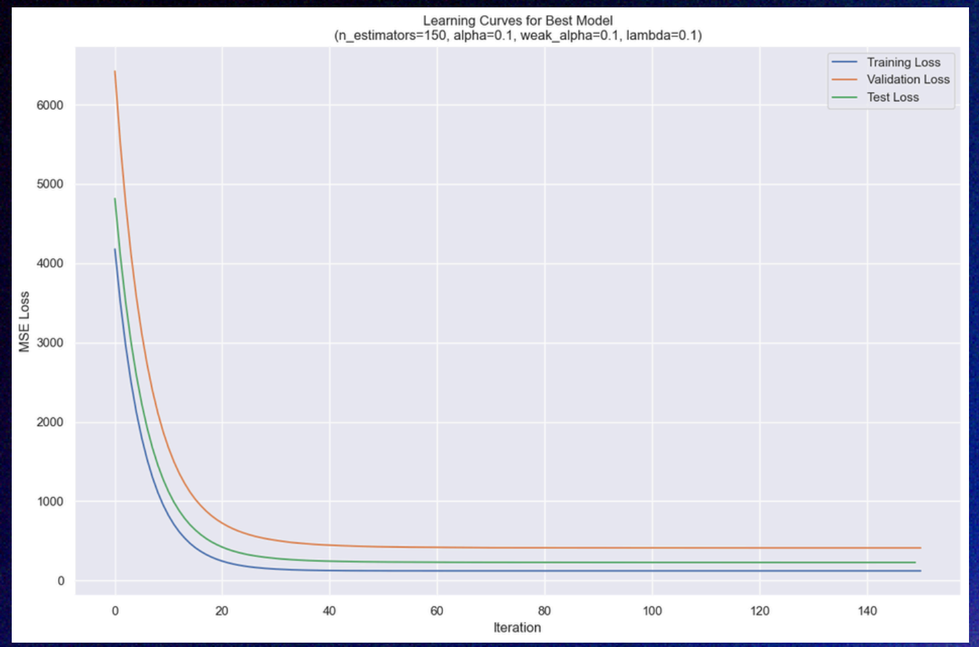


MSE Loss is 227.90
MAE Loss is 9.15

we see that training below testing again but loss are little improve from ensemble all the good gradient classifier to our H

Parameter Search and Analyst

- Next we will try to visualize the best prarameter
- Using search combination of paremeter approach
- Then Analyst the search value using pair graph

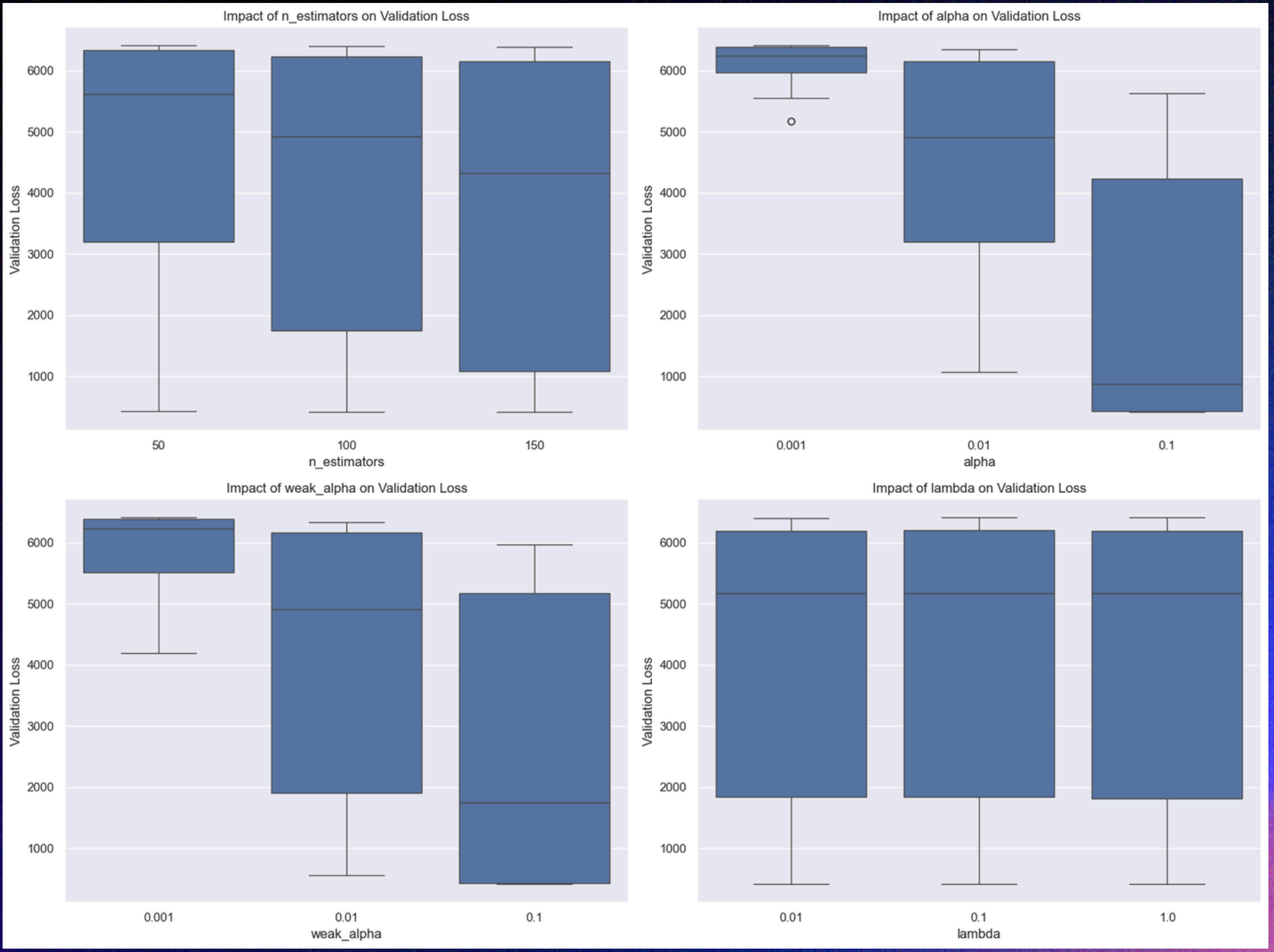


Best parameters:

- n_estimators: 150
- alpha: 0.1
- weak_alpha: 0.1
- lambda: 0.1
- MSE_test: 227.427263
- MAE_test: 9.165042

This still the same?

Parameter Search and Analyst cont.



n_estimators vs validation loss

- box plot indicate 90% data
- more n is better loss

alpha (ensemble) vs validation loss

- box plot indicate 90% data
- more is better : 0.1 is best learning rate for ensemble

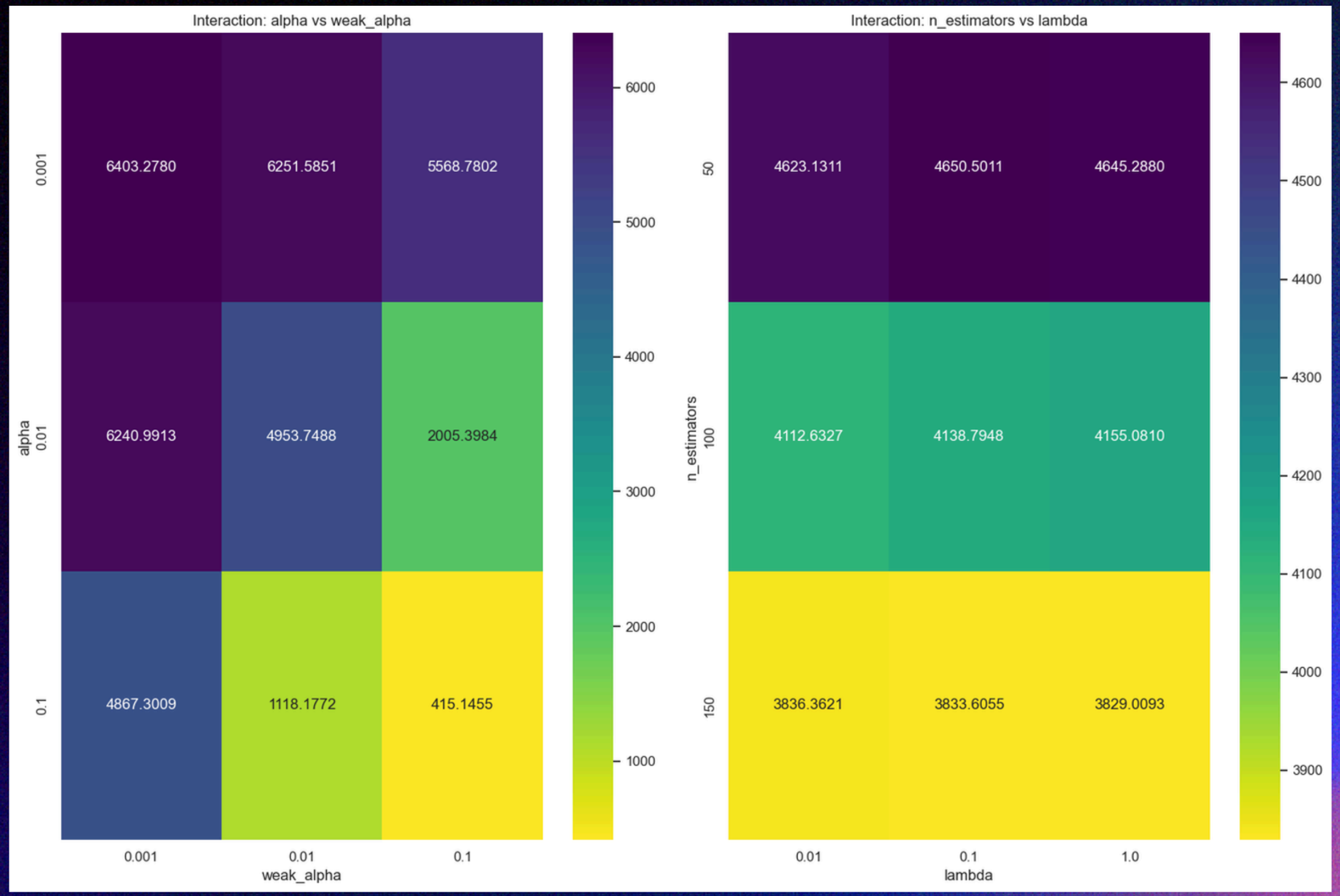
weak_alpha (estimators) vs validation loss

- box plot indicate 90% data
- more is better : 0.1 is best learning rate for all estimator

lambda (regularization) vs validation loss

- box plot indicate 90% data
- not effect

Parameter Search and Analyst cont.



weak_alpha vs alpha

- have linear correlation
- more value lead to make algorithm learn faster
- more is better

lambda vs n_estimators

- lambda not effect
- more n_estimators lead to better validation loss

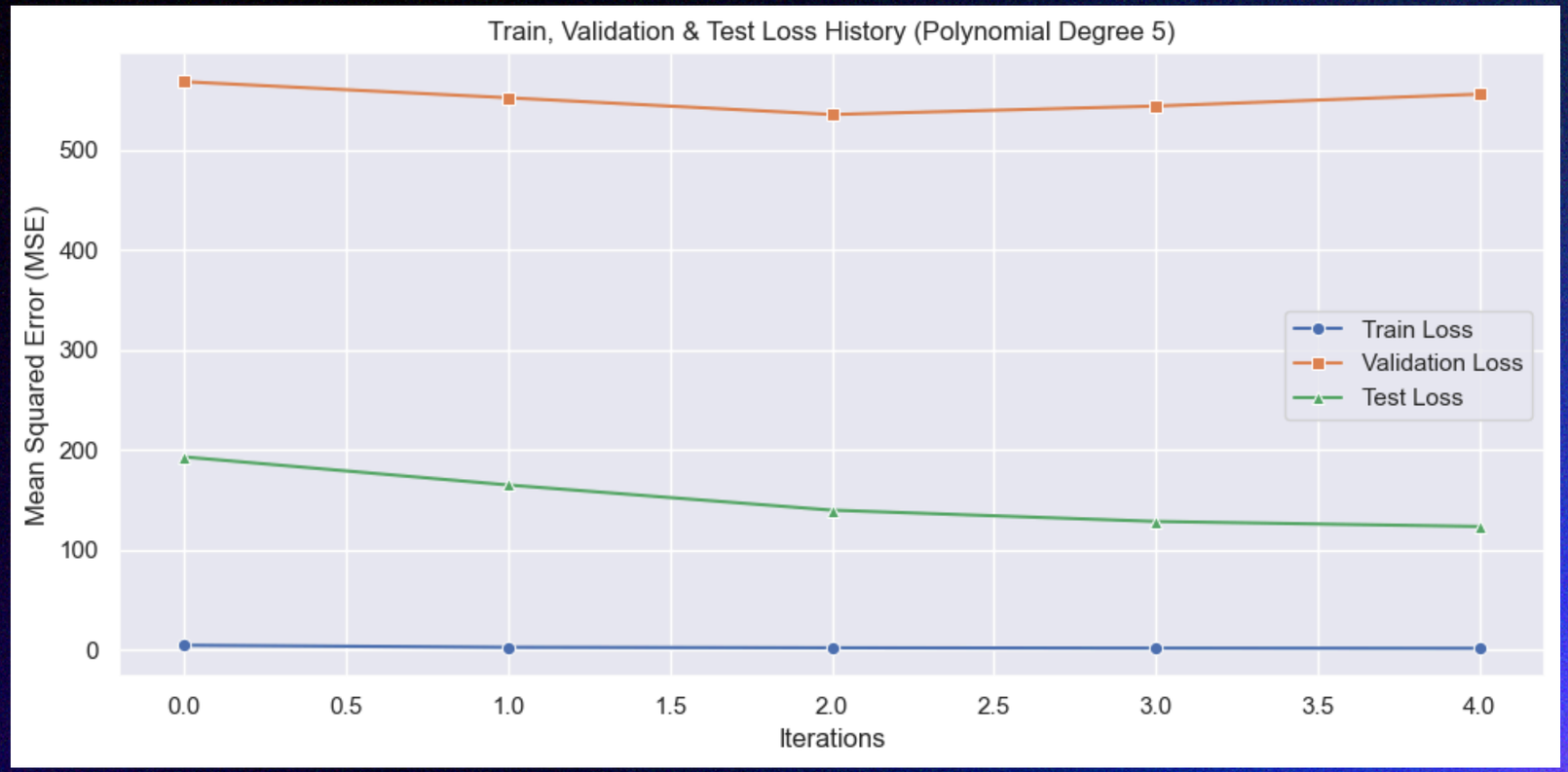
Kernelization

- As we see that we try to solve the variance and bias problem and It still not working for our dataset
- The fact of linear regression, It’s linear decision boundary
- So expand dimension may improve!
- Try using the Kernelization with Polynomail Degree 5 to solve this!
- Training and tuning parameter
- Evaluation

Kernelization cont.

- Step 1 expand dimension to
  - X_train
  - X_validation
  - X_test
  - by using sklearn
    - StandardScaler
    - PolynomialFeatures
- Step 2 train it as normal

Kernelization cont.



Min MSE is 123.29
Min MAE is 3.62

Yes - The kernelization that expand the feature dimension and solve the linear dicision boundary problem!

Thank You!

6610402230 Sirisuk Tharntham Section 1