Orange quality Classification

Machine Learning Final Presentation
- Yirong Wang and Helen Yuan

Introduction

Orange Quality Dataset

- 10 input features
- 1 output feature(orange quality: 1-5)
- 241 samples
- No Missing data

Binary Classification Problem

- Consider quality >=4 as **good** orange(worth to buy)

Data Processing

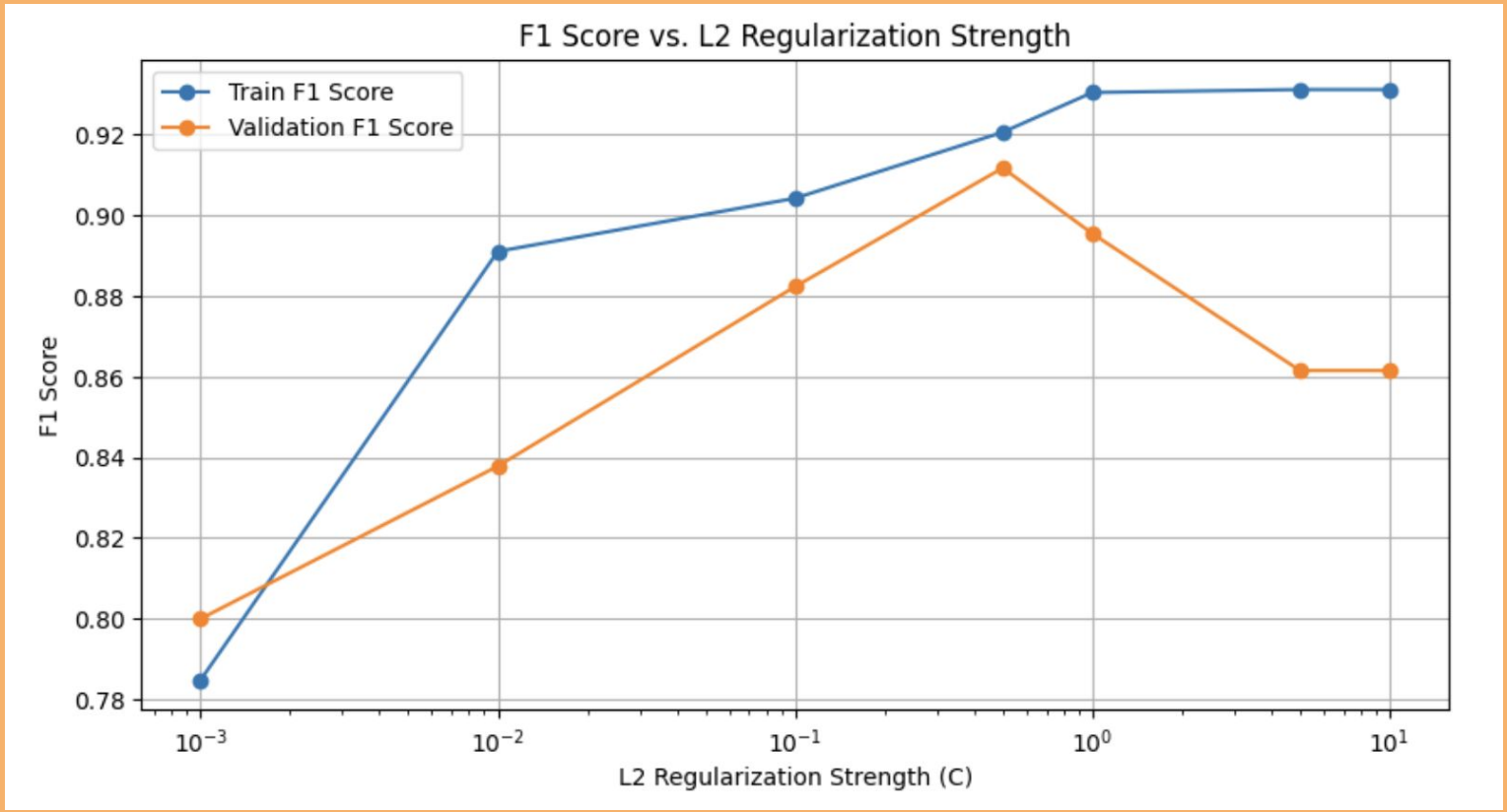
size	weight	sweetness	ph	softness	harvest_days	ripeness	color	variety	blemishes	quality
7.5	180	12	3.2	2	10	4	Orange	Valencia	N	4
8.2	220	10.5	3.4	3	14	4.5	Deep Orange	Navel	N	4.5
6.8	150	14	3	1	7	5	Light Orange	Cara Cara	N	5
9	250	8.5	3.8	4	21	3.5	Orange-Red	Blood Orange	N	3.5
8.5	210	11.5	3.3	2.5	12	5	Orange	Hamlin	Y (Minor)	4.5
6.7	126	9.1	3	2	25	2	Orange	Navel	N	1
7.2	160	9	3.5	3.5	9	4	Yellow-Orange	Tangelo (Hybrid)	N	4
6.5	130	13.5	2.8	1.5	5	4.5	Light Orange	Murcott (Hybrid)	N	4.5
8.8	240	7.5	4	5	18	3	Deep Orange	Moro (Blood)	Y (Sunburn)	3
7.8	190	12	3.1	2	11	4.5	Orange	Jaffa	N	5
9.5	270	6	4.2	4.5	24	2.5	Orange-Red	Cara Cara	Y (Mold Spot)	2.5
7.8	183	14.8	3.7	2	12	3	Deep Orange	Valencia	Y (Mold Spot)	4
8	200	10	3.5	3	13	4	Orange	Clementine	N	4.5
7	140	11	3.2	2.5	8	4.5	Deep Orange	Washington Navel	N	5
9.2	260	9.5	3.7	4.5	20	4	Orange-Red	Star Ruby	N	4
6.3	120	14.5	2.9	1	6	5	Light Orange	Tangerine	N	5
8.7	230	8	3.9	3.5	17	3.5	Orange	Ambiance	Y (Bruise)	3.5
9.6	218	14.1	4.2	4	11	1	Deep Orange	Cara Cara	Y (Sunburn)	4
7.5	247	9.1	3.3	4	24	5	Light Orange	Clementine	N	2
7.4	170	12.5	3	2	10	4	Yellow-Orange	Jaffa	N	4.5
10	300	7	4.1	5	25	3	Orange-Red	Blood Orange	N	3
8.1	205	11	3.4	2.5	14	4.5	Deep Orange	Murcott (Hybrid)	N	5
7.6	180	9	3.3	3	11	4	Orange	California Valencia	N	4.5
9.8	280	6.5	4.3	5	23	2.5	Orange-Red	Moro (Blood)	Y (Split Skin)	2
7.9	190	10.5	3.1	2.5	12	4	Orange	Honey Tangerine	N	4.5

Data Processing

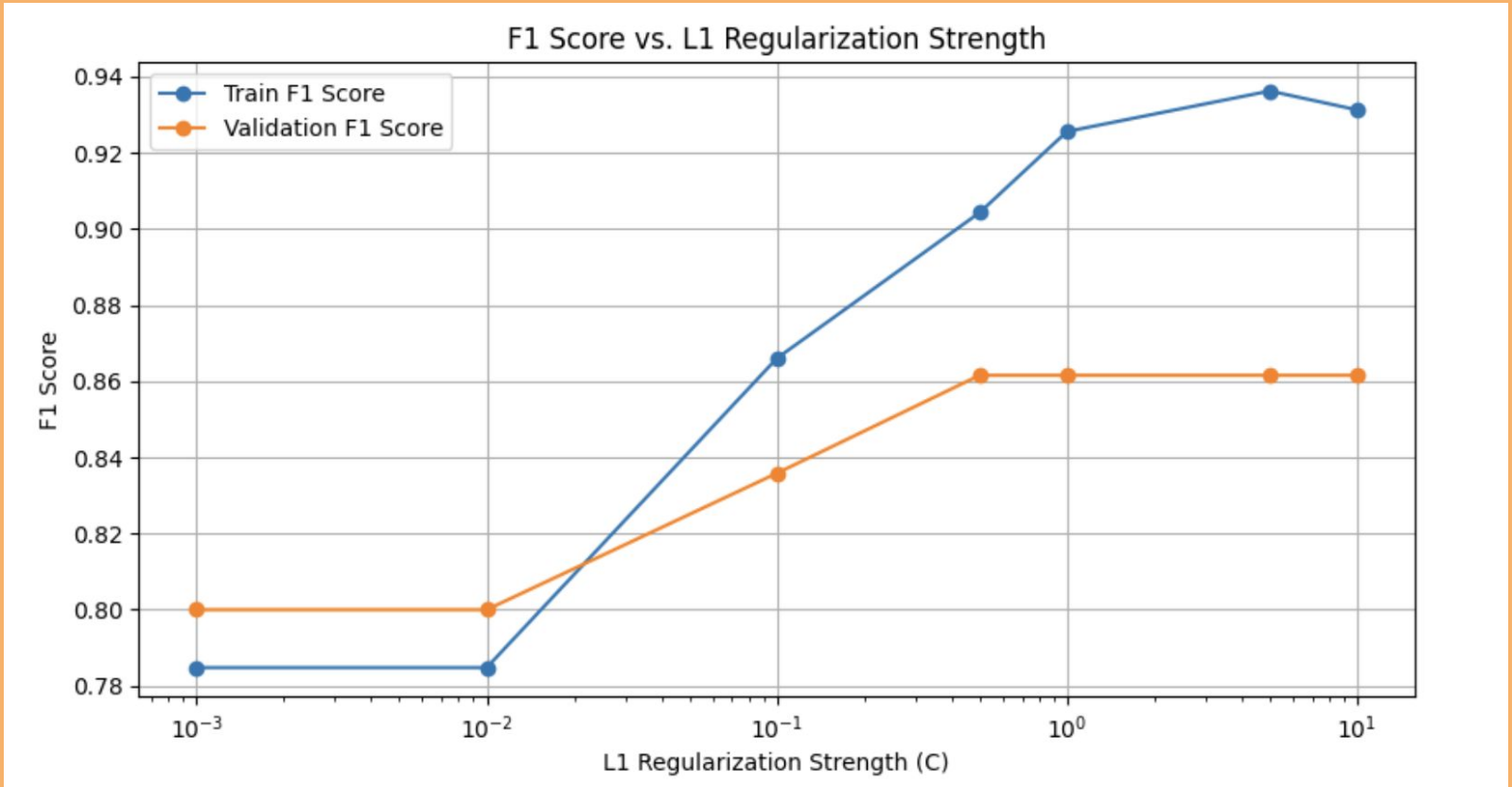
- One Hot Encoding (color, variety, blemishes)
  - 10 columns to 48 columns
- Output Feature Binarization
- Data splits
  - 60% train
  - 20% validation
  - 20% test

Logistic Regression

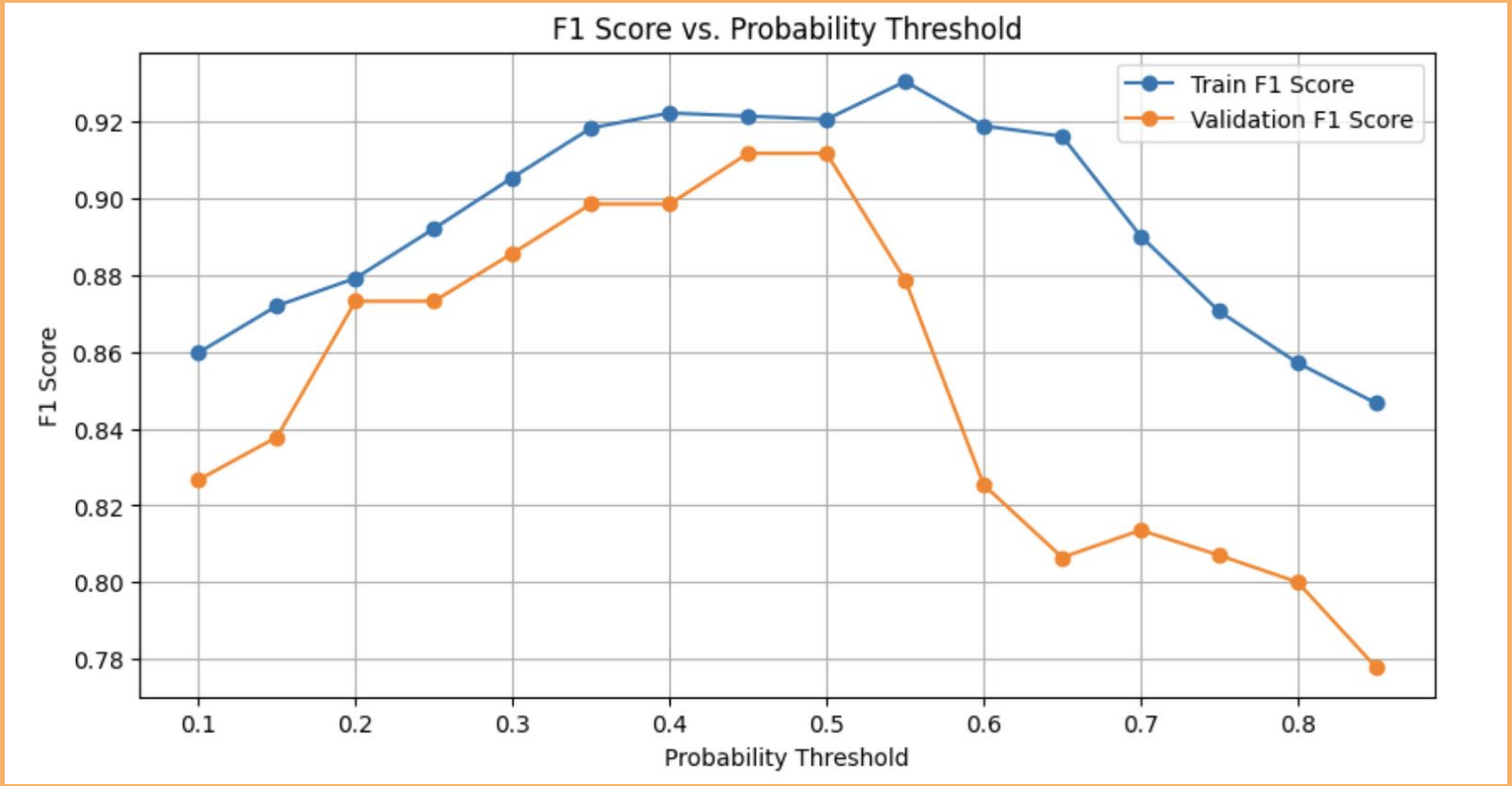
L2 Regularization– best strength Value 0.5



L1 Regularization– best strength Value 5



Different Probability Threshold – best 0.5



Feature Transformation: Second Degree

L2 Regularization + strength 0.5 + Threshold 0.5

With transformation ✅

- Training F1 Score: 0.9946524064171123
- Validation F1 Score: 0.9253731343283582

Without transformation

- Training F1 Score: 0.9206349206349206
- Validation F1 Score: 0.911764705882353

Test with Best Parameters

L2 Regularization

+

strength 0.5

+

Feature Transformation

+

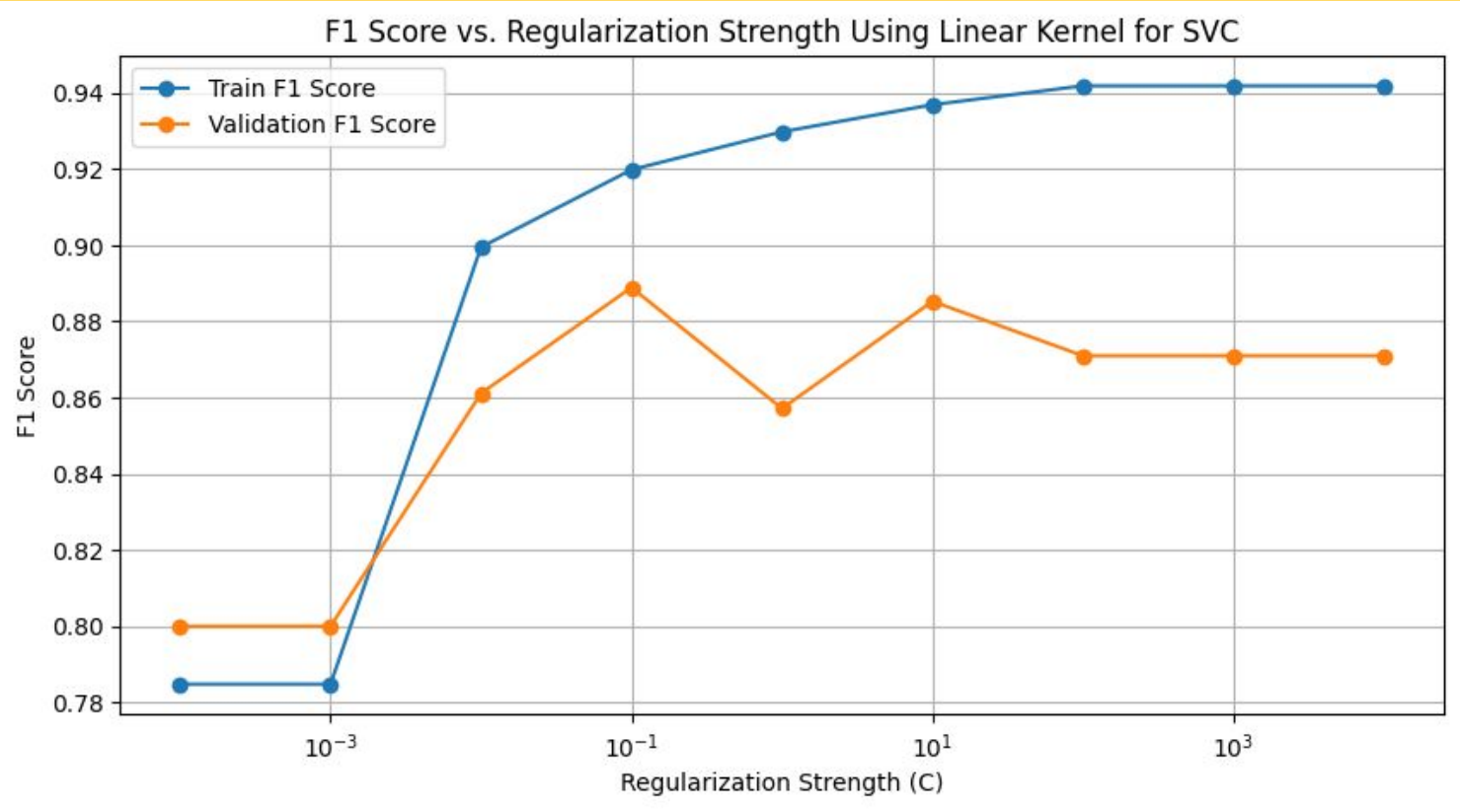
Probability Threshold 0.5

=

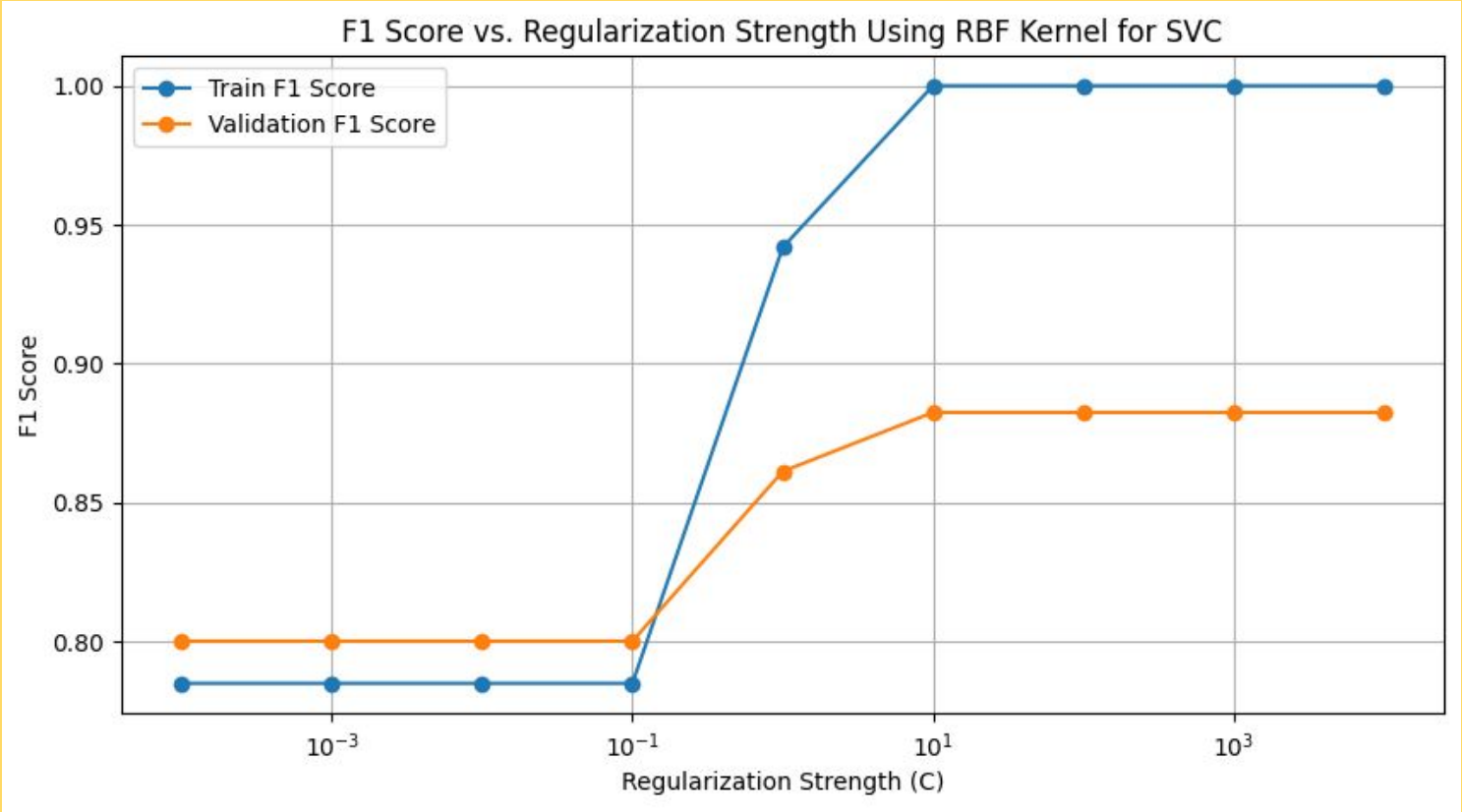
Test Set F1 Score: 0.84375

SVM

Regularization on Linear Kernel - Best strength Value 0.1



Regularization on RBF Kernel - Best strength Value 10



Test with Best Parameters

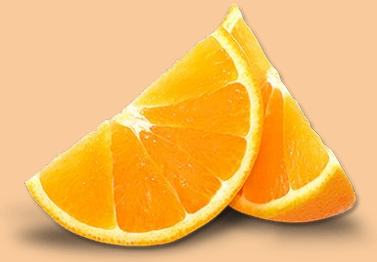
Linear Kernel

+

strength 0.1

=

Test Set F1 Score: 0.923076923076923



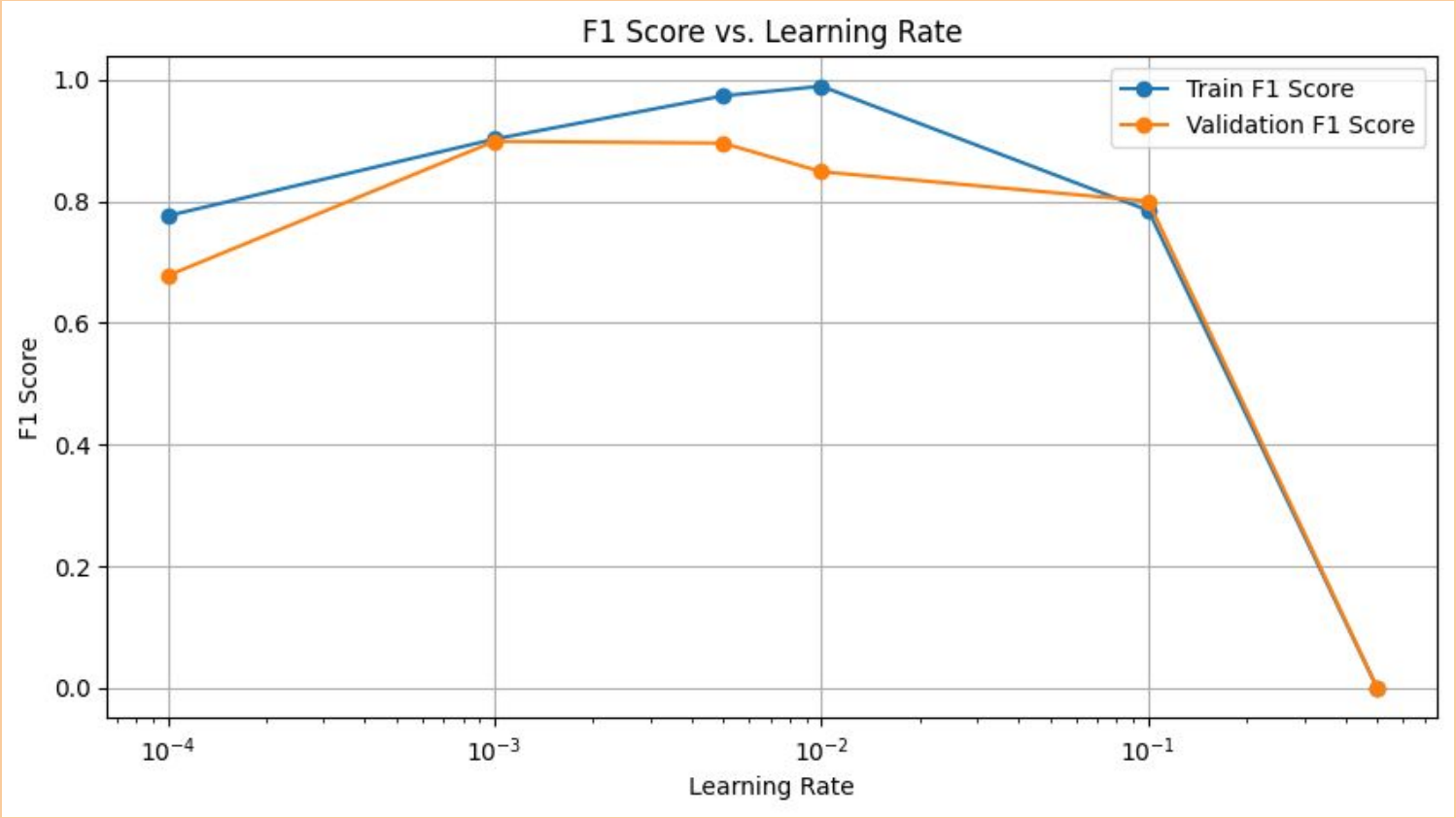
Neural network

7 Layer Neural Network

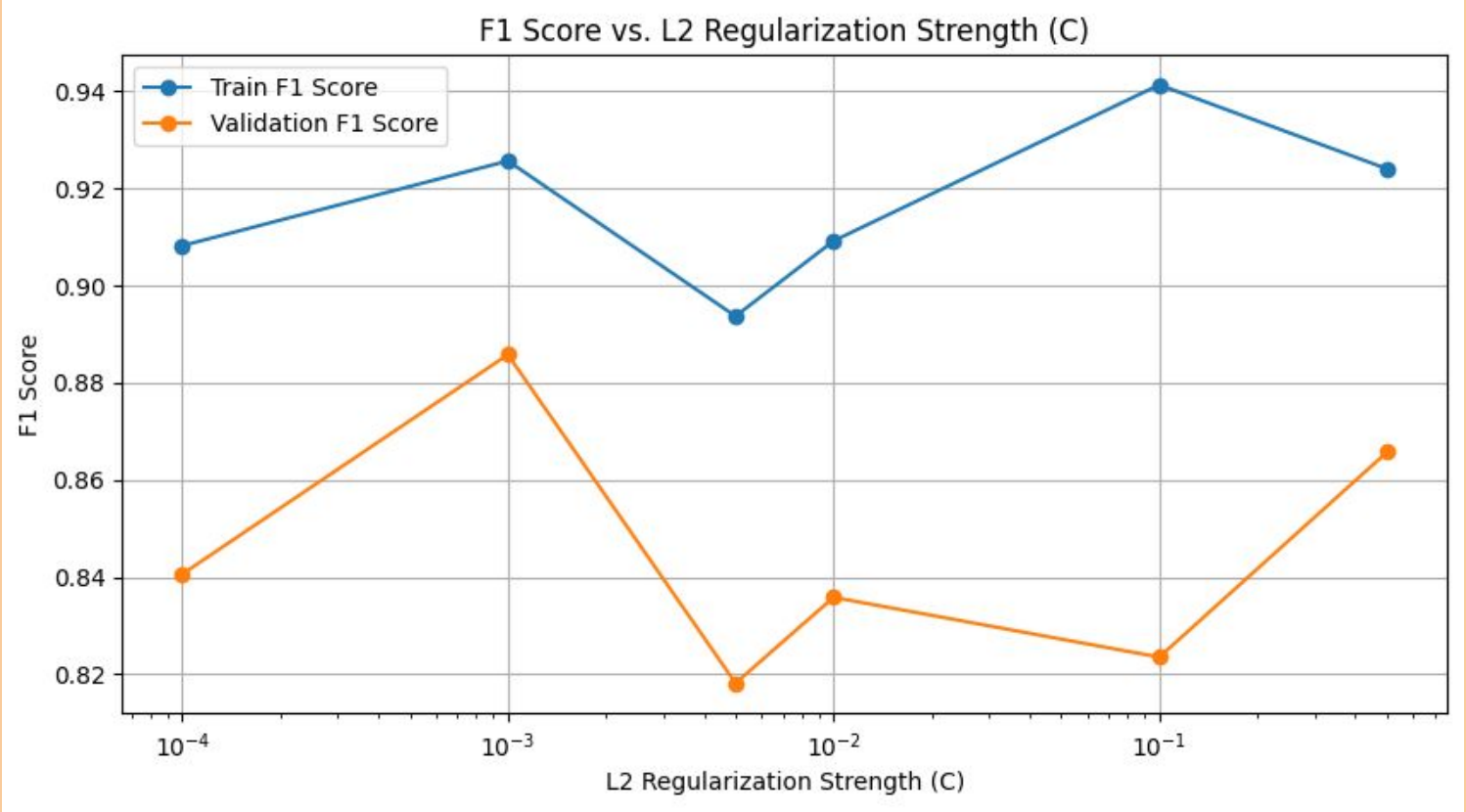
- ELU Activation Function for first 6 layers.
- Sigmoid Activation Function for the last layer.
- 144 -> 128 -> 64 -> 32 -> 16 -> 8 -> 1

```
model = tf.keras.Sequential([
  layers.Dense(128, activation='elu', input_shape=(144,), kernel_regularizer=l2(0.001)),
  layers.Dense(64, activation='elu', kernel_regularizer=l2(0.001)),
  layers.Dense(32, activation='elu', kernel_regularizer=l2(0.001)),
  layers.Dense(16, activation='elu', kernel_regularizer=l2(0.001)),
  layers.Dense(8, activation='elu', kernel_regularizer=l2(0.001)),
  layers.Dense(1, activation='sigmoid', kernel_regularizer=l2(0.001)),
])
```

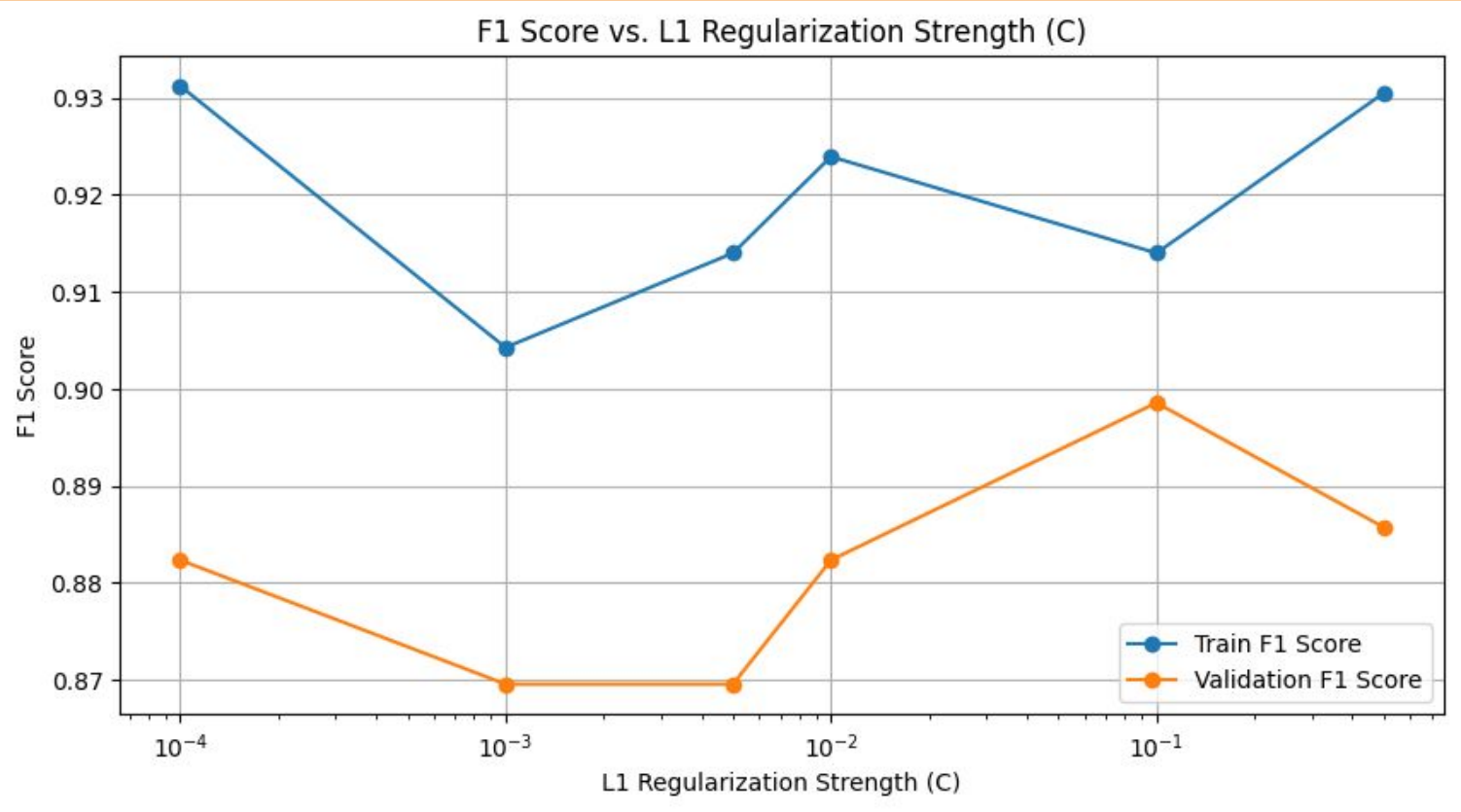
Best Learning Rate - 0.001



L2 Regularization- best strength Value 0.001



L1 Regularization– best strength Value 0.1



Feature Transformation: Second Degree

L1 Regularization + strength 0.1 + Learning Rate 0.001

With transformation

- Training F1 Score: 0.9304812834224598
- Validation F1 Score: 0.8169014084507042

Without transformation ✅

- Training F1 Score: 0.9247311827956989
- Validation F1 Score: 0.8695652173913043

Test with Best Parameters

L1 Regularization

+

strength 0.1

+

Learning Rate 0.001

=

Test Set F1 Score: 0.9253731343283582

Conclusion

Logistic Regression: 0.84375

SVM: 0.923076923076923

Neural Network: 0.9253731343283582

Thank You!!

Machine Learning Final Presentation
- Yirong Wang and Helen Yuan