Machine Learning 2024 Spring
HW1: Linear Regression
109511286 蔡佩蓉

Introduction

Linear regression is a fundamental technique in machine learning used to model the relationship between a set of input features and a target variable. In this report, we explore the application of linear regression with basis functions and regularization techniques to predict the popularity of songs based on various features. We aim to minimize the error between the predicted popularity and the actual popularity of songs.

Methodology

Basis Functions

We employ logistic sigmoid basis functions to capture nonlinear relationships between input features and the target variable. The basis functions are defined as:

$$\phi_{k,j}\,(x_k) = \begin{cases} 1, & j = 0 \\ \sigma\left(\frac{x_k - \mu_j}{s}\right), & j = 1, \ldots, M-1 \end{cases}$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the logistic sigmoid function, $s = 0.1$, and μ_j is computed as:

$$\mu_j = \frac{3\left(-M + 1 + 2(j-1)\frac{M-1}{M-2}\right)}{M}$$

for $j = 1, \ldots, (M-1)$, where M is the order of the basis functions.

These basis functions transform the input features into a higher-dimensional space, allowing the model to capture complex patterns.

Model Training

We trained linear regression models using the training data by minimizing the error function, which is the sum of squared differences between the predicted and actual target values. Additionally, we introduced regularization to prevent overfitting by adding a penalty term to the error function.

The error function $E(w)$ without regularization is given by:

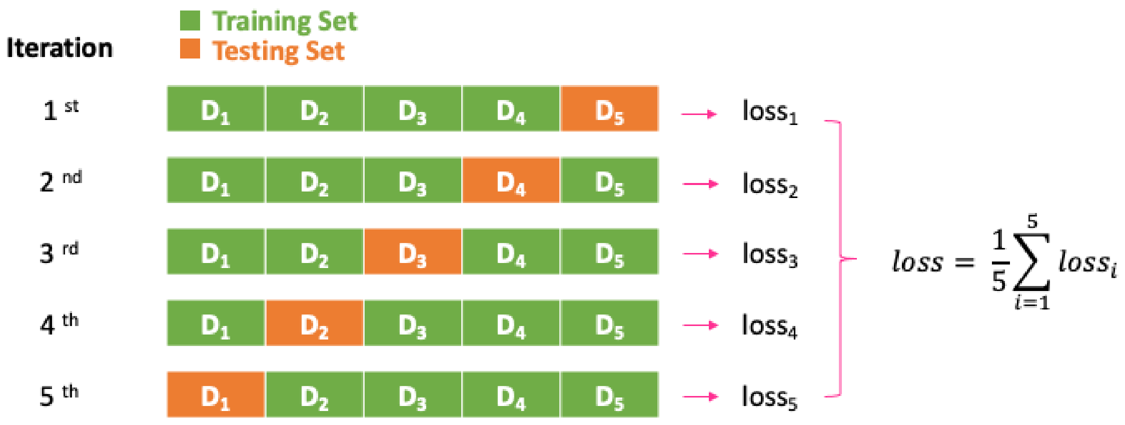
$$E(w) = \frac{1}{2}\sum_{i=1}^{N}\left\{y\left(x_{i,1}, \ldots, x_{i,K}, w\right) - t_i\right\}^2$$

where $y(x_1, \ldots, x_K, w) = \sum_{k=1}^{K}\sum_{j=0}^{M-1} w_{k,j}\phi_{k,j}(x_k)$ and $\phi_{k,j}(x_k)$ are the basis functions

Cross-Validation

To select the optimal order of basis functions (M), we employ 5-fold cross-validation. We partition the training data into K equal parts ($K = 5$ in this Homework), then take turns to use one of them as the validation set (indicated as testing set in the illustration below), and then use the remaining $K - 1$ parts as the training set to train the model.

We repeat this process for different values of M and choose the one that minimizes the average validation error. This is because the results of training on different subsets are averaged to reduce variance, so the performance of the model will not be so sensitive to the partitioning of the data.



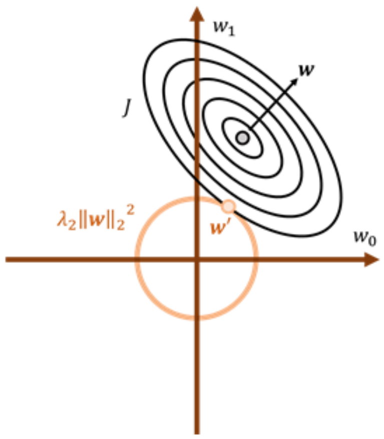
Regularization

To prevent overfitting, we incorporate regularization into the model by adding a penalty term to the error function:

$$\tilde{E}(w) = \frac{1}{2}\sum_{i=1}^{N}\{y\left(x_{i,1},\dots,x_{i,K},w\right) - \ t_i\}^2 \ + \frac{\lambda}{2}\,\|w\|^2$$

where λ is the regularization parameter, and $\|w\|^2 = \sum_{k=1}^{K}\sum_{j=0}^{M-1} w_{k,j}^2$ is the L2 norm of the weight vector.

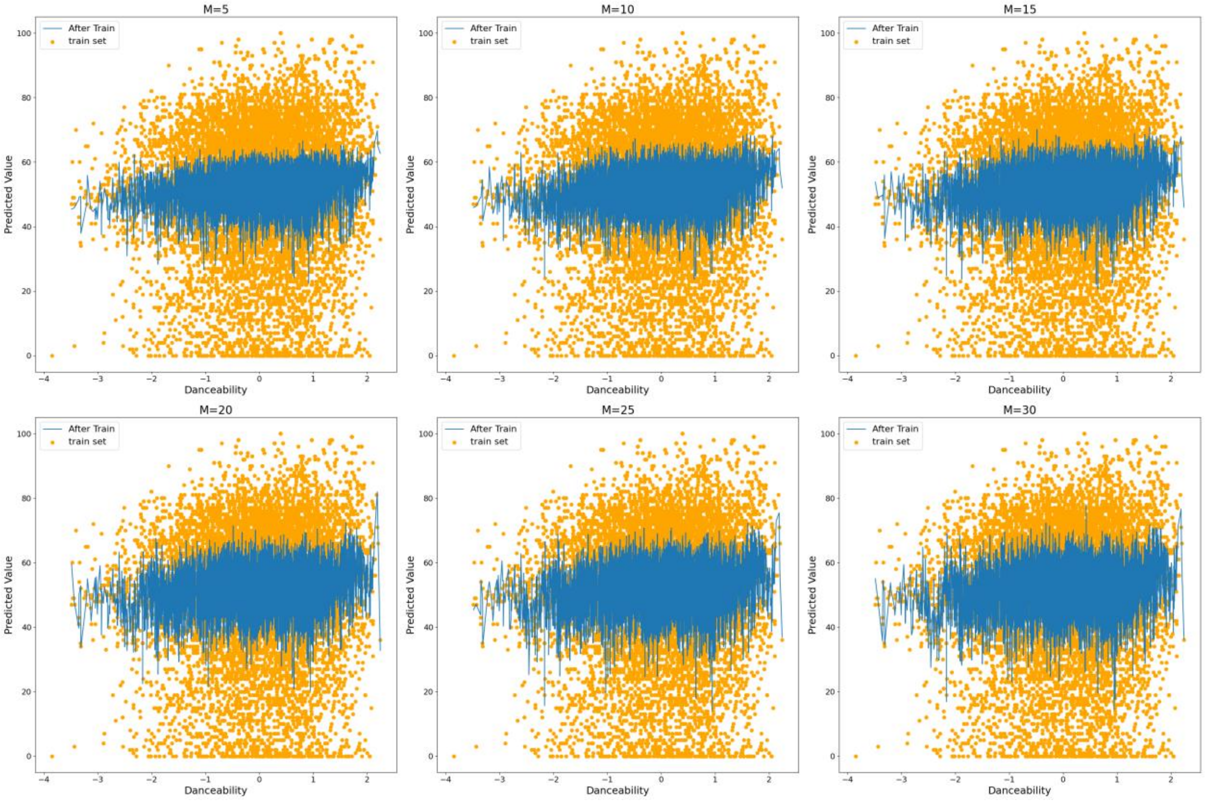
The solution we get when not using regularization is w (overfitting case). When we use L2 norm restrictions, the solution we get will be w', overfitting reduced.



Results

Fitting Curve

Plotting Fitting Curve for the third input feature (x_3: danceability) with M = 5, 10, 15, 20, 25, 30.



As M increases, the curves become more flexible and better capture the underlying patterns in the data (by observing the amplitude).

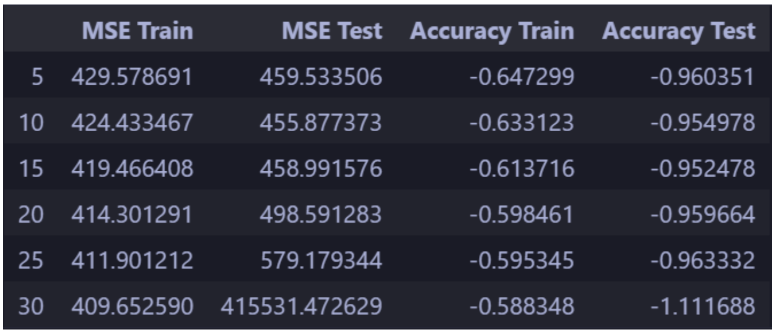
Mean Square Error and Accuracy Evaluation

We plotted the mean square error and accuracy for different values of M on both the training and testing sets. The accuracy of the training set and the testing set are evaluated as follows:

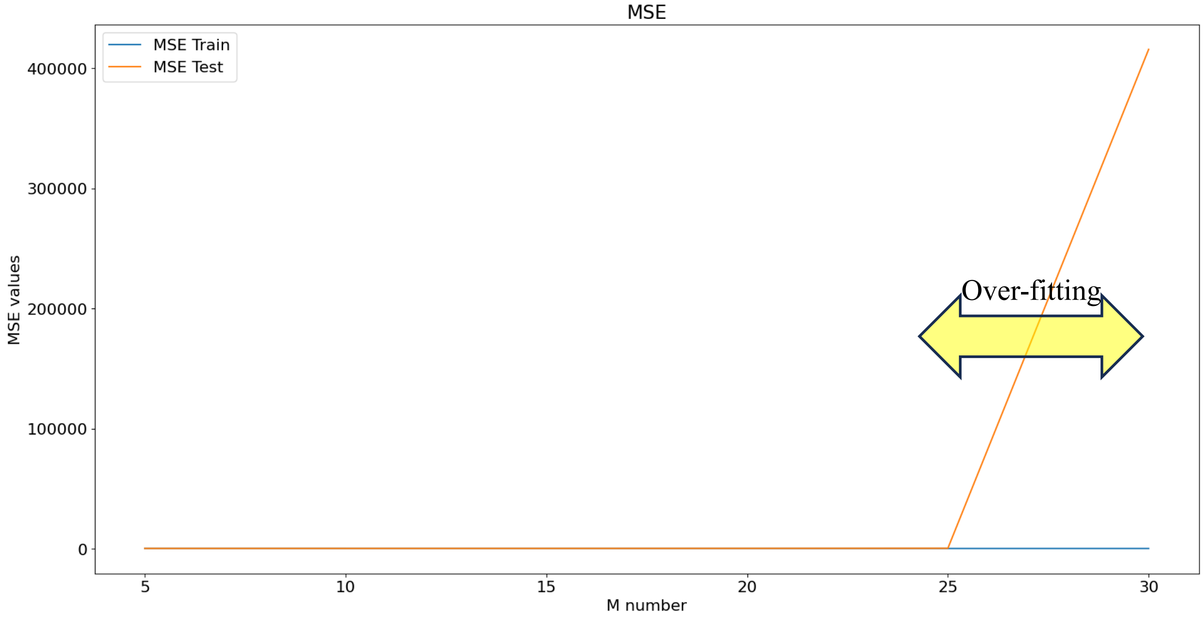
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - t_i)^2$$

$$accuracy = 1 - \frac{1}{N_d}\sum_{i=1}^{N_d}\left|\frac{y - t_i}{t_i}\right|$$

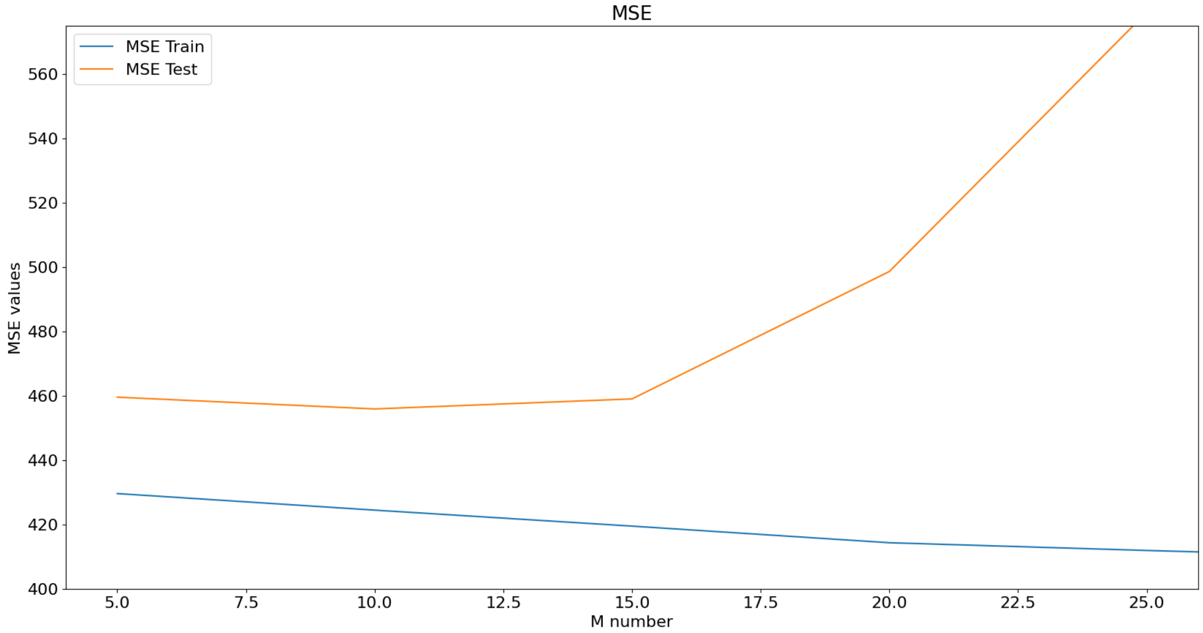
- N_d: number of data points in a data set
- Only for evaluating the accuracy, if the target value $t_i = 0$, replace the denominator as 1.



	MSE Train	MSE Test	Accuracy Train	Accuracy Test
5	429.578691	459.533506	-0.647299	-0.960351
10	424.433467	455.877373	-0.633123	-0.954978
15	419.466408	458.991576	-0.613716	-0.952478
20	414.301291	498.591283	-0.598461	-0.959664
25	411.901212	579.179344	-0.595345	-0.963332
30	409.652590	415531.472629	-0.588348	-1.111688



Magnified plot:

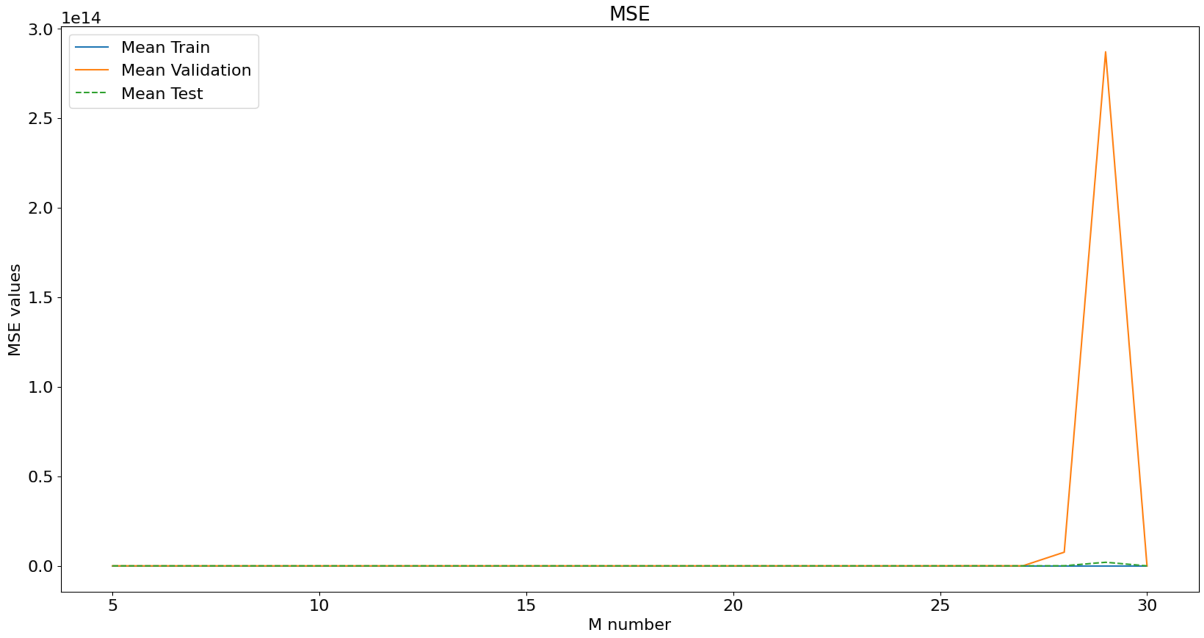


As expected, the training error decreases with increasing M, but the testing error initially decrease and then start to increase due to overfitting. The accuracy follows a similar trend. Besides, we observed that no matter which M is chosen, the MSE of testing data is always larger than that of training data.

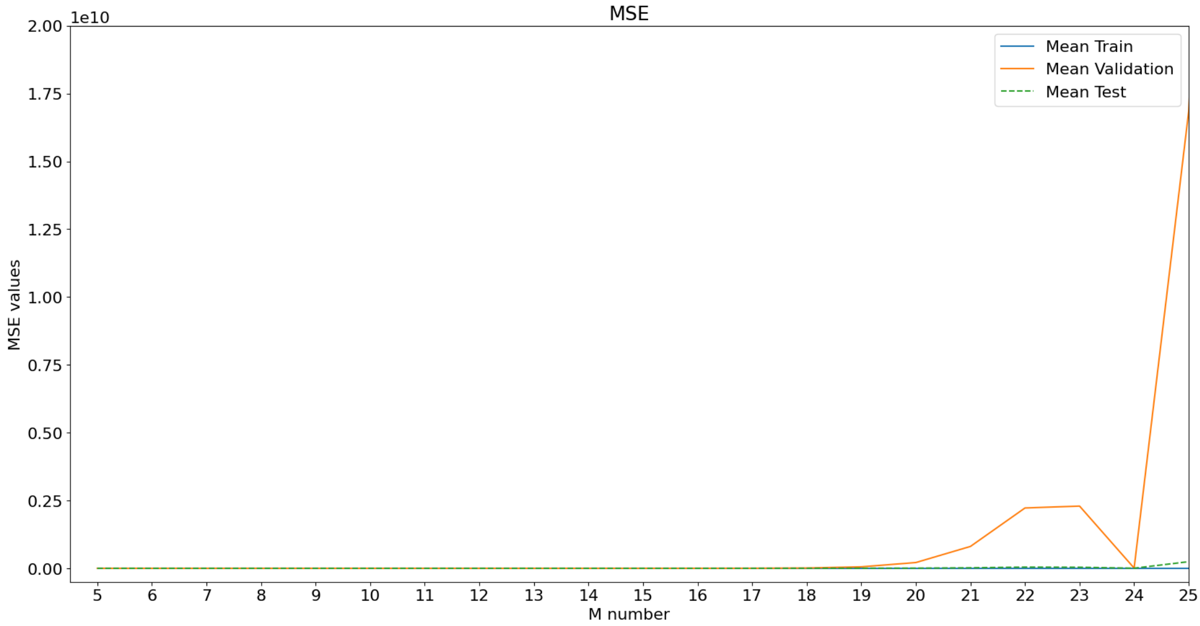
K-Fold Cross-Validation

We applied 5-fold cross-validation to select the best value of M. After evaluating the model performance on the validation sets, we chose the value of M that yielded the lowest average validation error.

	MSE Mean Validation
5	457.1496
6	481.8451
7	462.0706
8	475.0567
9	489.3098
10	978.4362
11	8699.563
12	787.8435
13	1751.646
14	1533.995
15	52270.51
16	143716.1
17	1101454
18	9027750
19	53815646
20	2.12E+08
21	8.06E+08
22	2.22E+09
23	2.29E+09
24	7700240
25	1.68E+10
26	5.56E+10
27	1.12E+10
28	7.64E+12
29	2.87E+14
30	1.5E+10

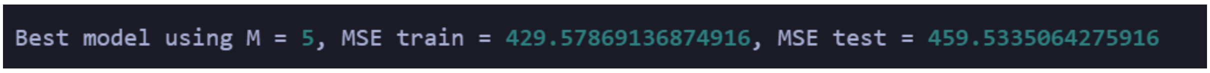


Magnified plot:

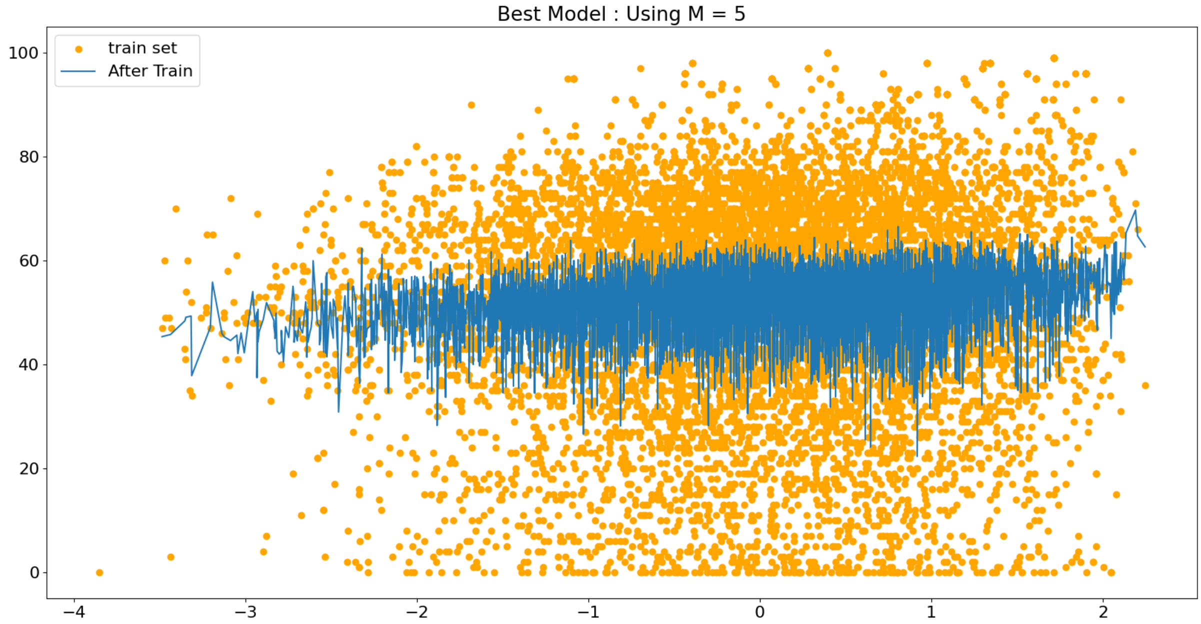


(Note that the green dashed line is the MSE of testing set and it should not be considered when we choose models)

Hence, we choose M = 5 as the best order M as it has smallest MSE for validation set. Then, I use the best order M and all the training data to find the fitting curve. The result is showed as follow.

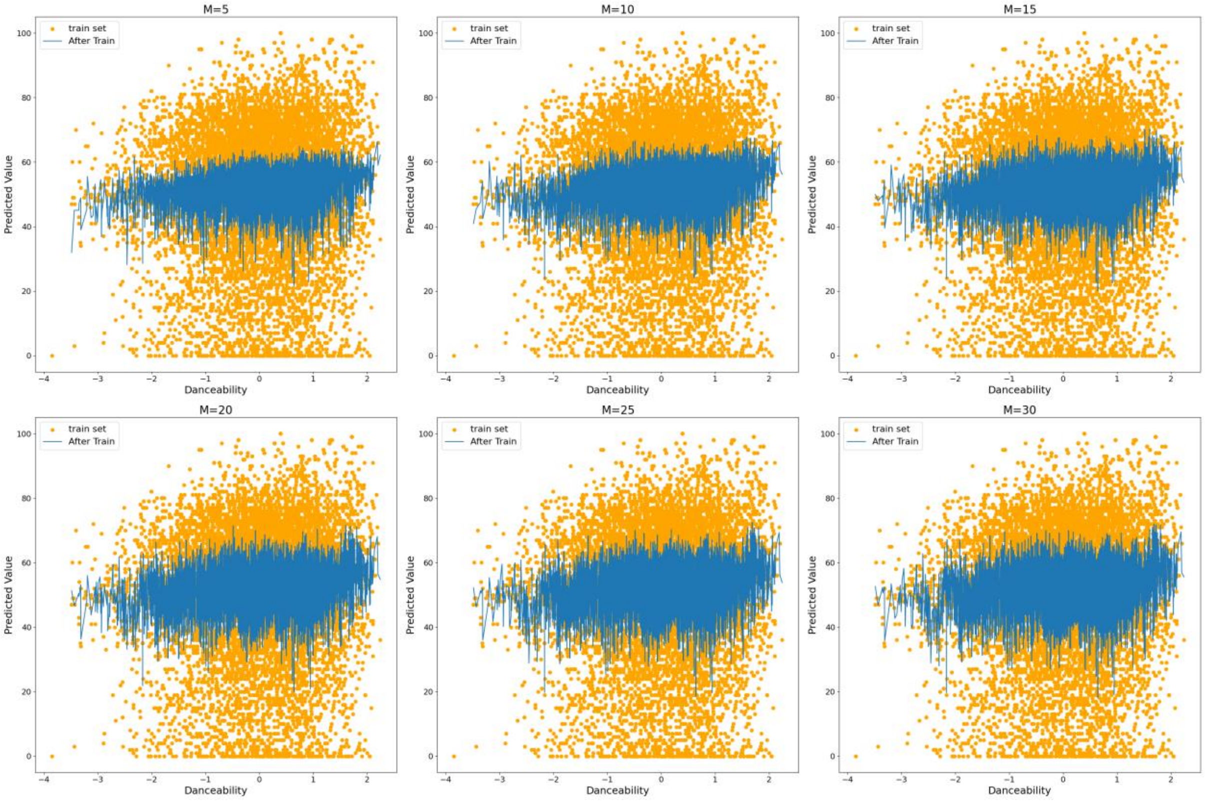


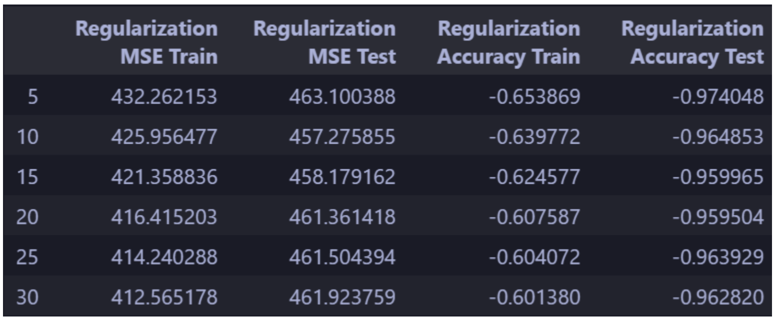
```
Best model using M = 5, MSE train = 429.57869136874916, MSE test = 459.5335064275916
```



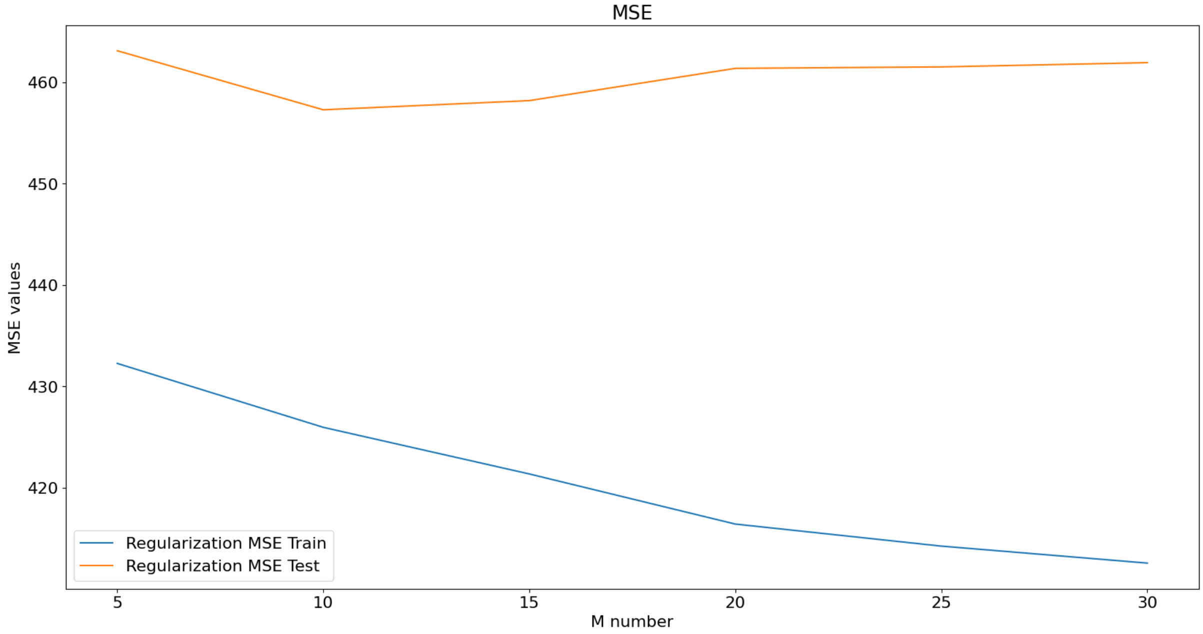
Regularization

We applied regularization to the model by introducing a penalty term in the error function. We evaluated the model's performance on the testing set with different values of the regularization parameter (λ) and observed its effect on mean square error.



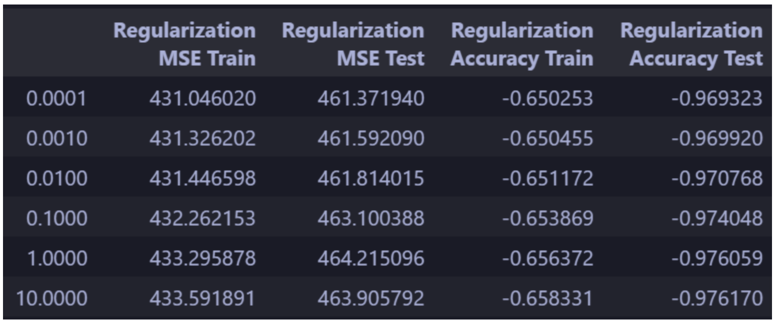


	Regularization MSE Train	Regularization MSE Test	Regularization Accuracy Train	Regularization Accuracy Test
5	432.262153	463.100388	-0.653869	-0.974048
10	425.956477	457.275855	-0.639772	-0.964853
15	421.358836	458.179162	-0.624577	-0.959965
20	416.415203	461.361418	-0.607587	-0.959504
25	414.240288	461.504394	-0.604072	-0.963929
30	412.565178	461.923759	-0.601380	-0.962820

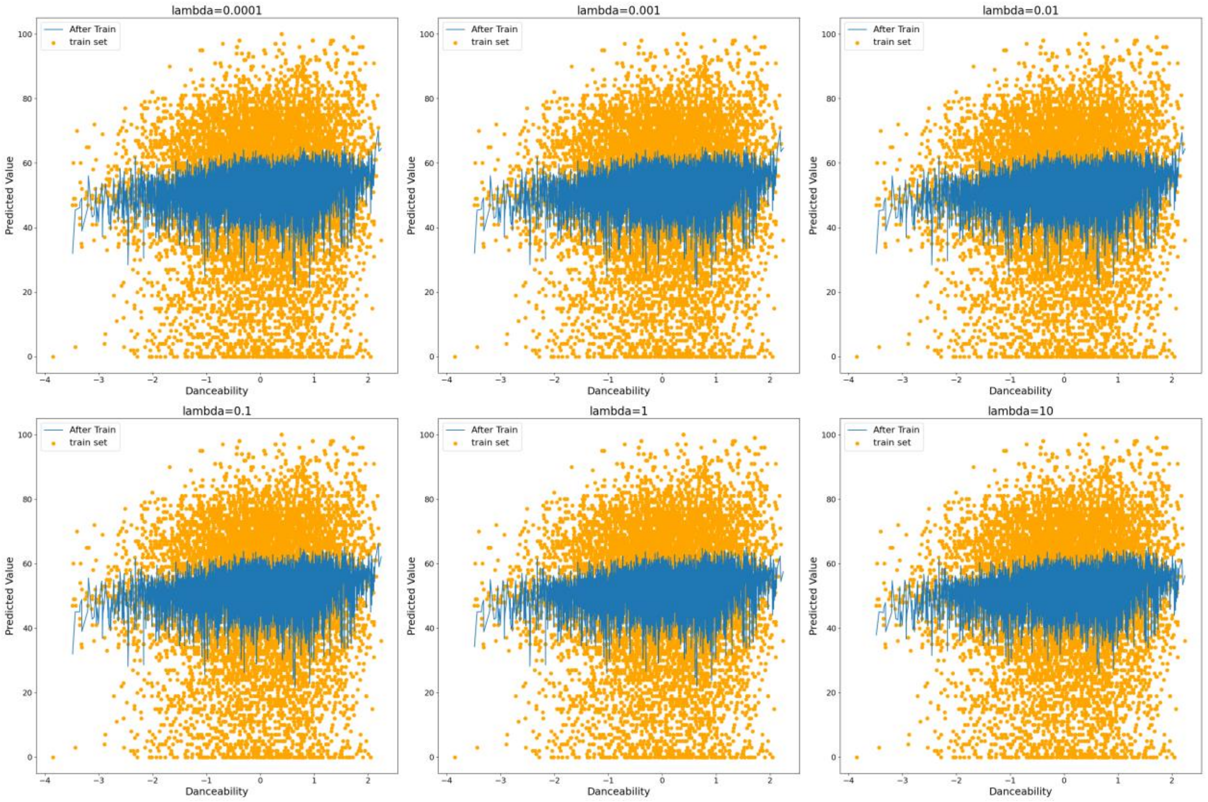


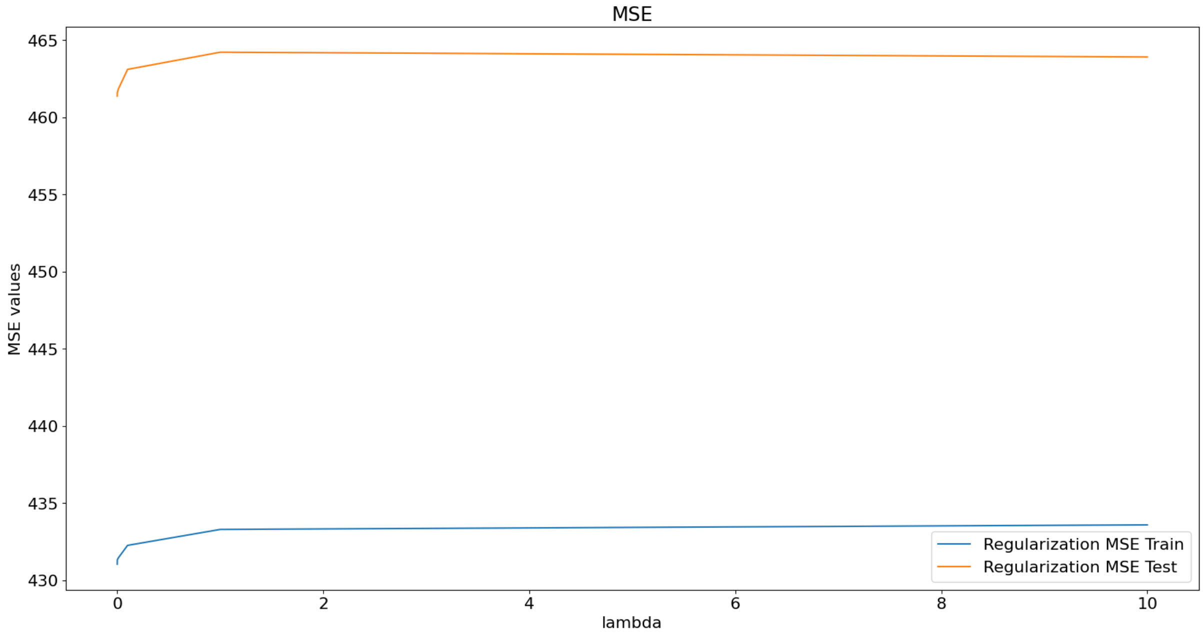
Comparing with the previous result, we observed that the addition of regularization term solves the overfitting problem and the generalization ability of the model also relatively improved. The MSE for testing data doesn't drastically increase for large order M and the accuracy of test set is significantly reduced. This is because some parameters become large when over-fitting occurs, and the regularization term $\frac{\lambda}{2}\|w\|^2$ in error function limits the value of these parameters.

We observed the effect on output with various λ values ($\lambda = 0.0001, 0.001, 0.01, 0.1, 1, 10$)



	Regularization MSE Train	Regularization MSE Test	Regularization Accuracy Train	Regularization Accuracy Test
0.0001	431.046020	461.371940	-0.650253	-0.969323
0.0010	431.326202	461.592090	-0.650455	-0.969920
0.0100	431.446598	461.814015	-0.651172	-0.970768
0.1000	432.262153	463.100388	-0.653869	-0.974048
1.0000	433.295878	464.215096	-0.656372	-0.976059
10.0000	433.591891	463.905792	-0.658331	-0.976170

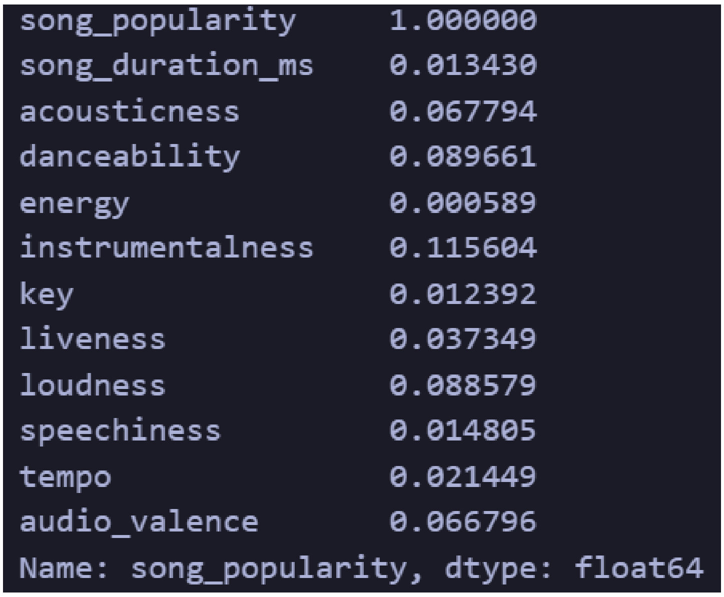




The lambda parameter controls the model elasticity (limit the size of weights so that the regression curve becomes less complex), if lambda becomes larger, it will make the model no elasticity or overfitting, otherwise, it becomes more elastic or overfitting.

Discussion

Which features do you consider the most important? Why do you consider your selected features to be the most important?



```
song_popularity        1.000000
song_duration_ms       0.013430
acousticness           0.067794
danceability           0.089661
energy                 0.000589
instrumentalness       0.115604
key                    0.012392
liveness               0.037349
loudness               0.088579
speechiness            0.014805
tempo                  0.021449
audio_valence          0.066796
Name: song_popularity, dtype: float64
```

Features with higher absolute correlation coefficients (with target variable) are usually more important. Hence, according to the absolute correlation coefficients shown above, we consider the feature, instrumentalness, is the most important.

Conclusion

In conclusion, linear regression with basis functions and regularization is a powerful technique for modeling complex relationships in data. By selecting appropriate basis functions and regularization parameters, we can build models that generalize well to unseen data. However, careful consideration must be given to avoid overfitting and underfitting. Cross-validation provides a robust method for model selection, while regularization helps prevent overfitting by penalizing overly complex models. Overall, this approach can be valuable in various applications where predictive modeling is required.