Machine Learning Project - Predicting Airfare Prices

Ruilin Zhou

April 29, 2024

Introduction

The airline industry is marked by dynamic pricing strategies.

Objective : Develop a predictive model for airline ticket prices.

Dataset : Flight data including historical prices, departure date, plane type, oil price, etc.

Dataset

Departure_Date	Departure_Day	Departure_Time	Plane_Type	Oil_Price	Days_In_Advance	Sale_Day	Price3	Price2	Price1	Ticket_Price
10/13/23	Friday	1:30pm	B737	82.92	57	Thursday	194	194	193	185
10/13/23	Friday	1:30pm	B737	84.24	56	Friday	194	193	185	186
10/13/23	Friday	1:30pm	B737	84.99	55	Saturday	193	185	186	194
10/13/23	Friday	1:30pm	B737	85.56	54	Sunday	185	186	194	192
10/13/23	Friday	1:30pm	B737	84.91	53	Monday	186	194	192	192

Departure_Date	Departure_Day	Departure_Time			Plane_Type			Oil_Price	Days_In_Advance	Sale_Day	Price3	Price2	Price1
13	4.0	13.5	0.0	0.0	1.0	0.0	0.0	82.92	57	3.0	194	194	193
13	4.0	13.5	0.0	0.0	1.0	0.0	0.0	84.24	56	4.0	194	193	185
13	4.0	13.5	0.0	0.0	1.0	0.0	0.0	84.99	55	5.0	193	185	186
13	4.0	13.5	0.0	0.0	1.0	0.0	0.0	85.56	54	6.0	185	186	194
13	4.0	13.5	0.0	0.0	1.0	0.0	0.0	84.91	53	0.0	186	194	192

Oil_Price : Closing price of Brent Crude oil on the day of ticket sale

Days_In_Advance : Number of days ahead of departure date

Sale_Day : Day of the week of ticket sale

Price1/2/3 : Price of ticket 1/2/3 days ago

Ticket_Price : Price of ticket for the specific flight (Target)

Datasetlink :https://www.kaggle.com/datasets/andrewrliu/airfare

Supervised Analysis

External library: NumPy, scikit-learn and Matplotlib.

Three models chosen: Linear Regression, Support Vector Regression, and Neural Networks.

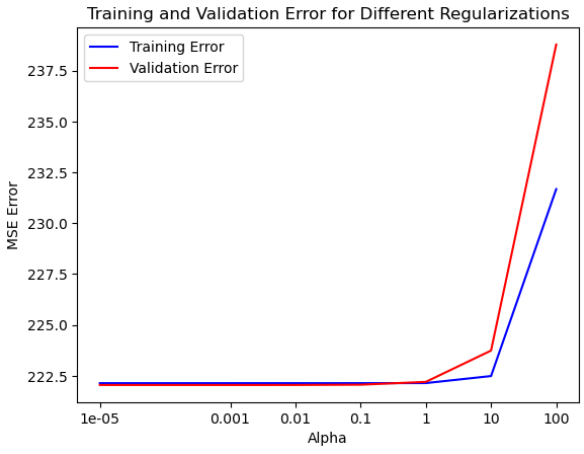
Optimal parameters: determined for each model based on validation set performance (Metric: MSE error).

Normalization: MinMax and Standardization with zero mean and unit variance.

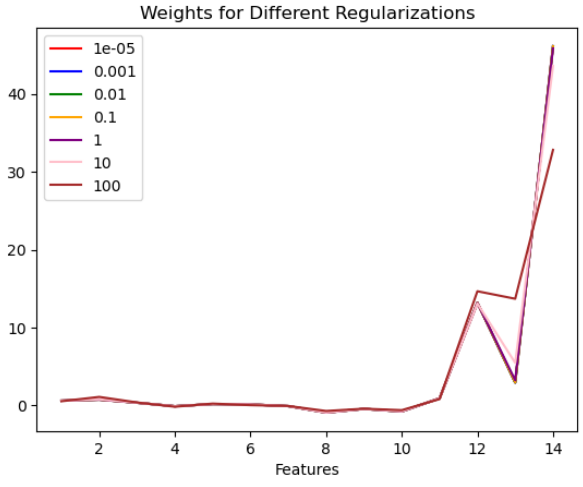
For SVR, ϵ values are chosen with consideration to MSE error from predicted results of linear regression model.

For each model, applied different regularization (Sampled in log scale), feature transformations (RBF, Polynomial Kernels) and parameters.

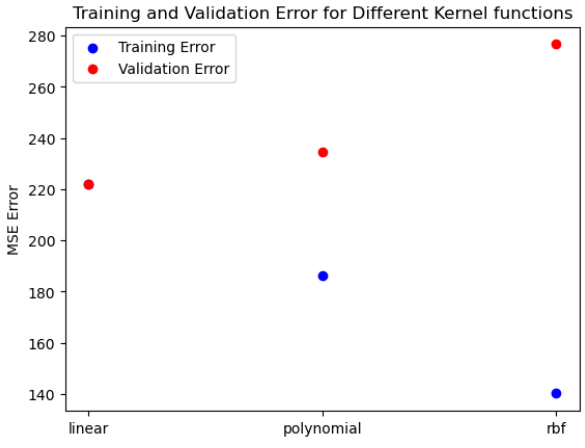
Linear regression - MSE error



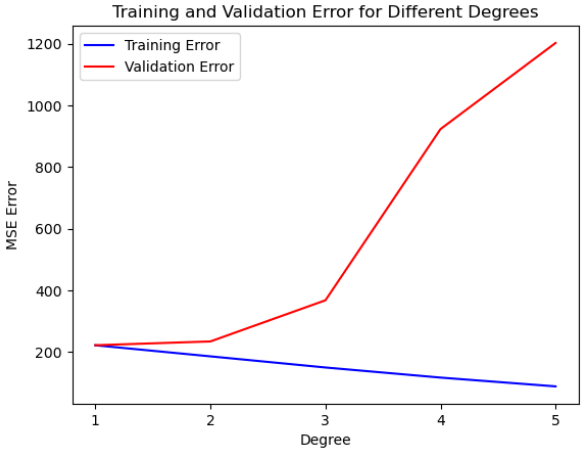
(a) Regularization



(b) Feature weights

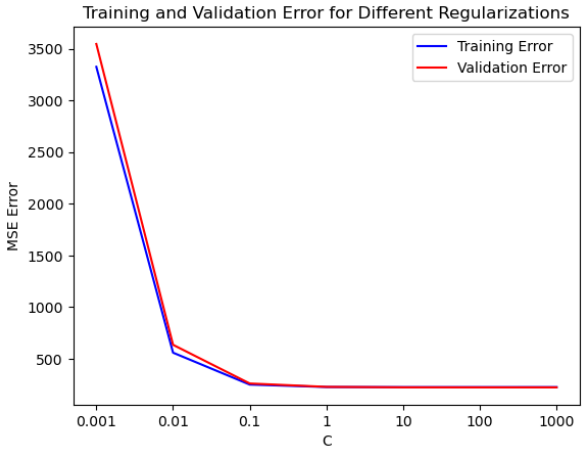


(c) Feature transformations

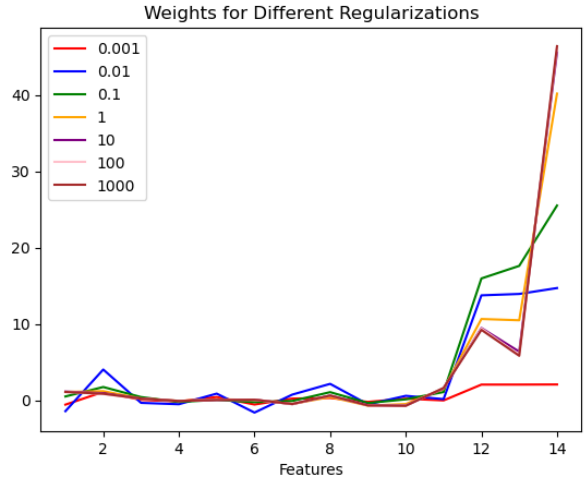


(d) Polynomial degrees

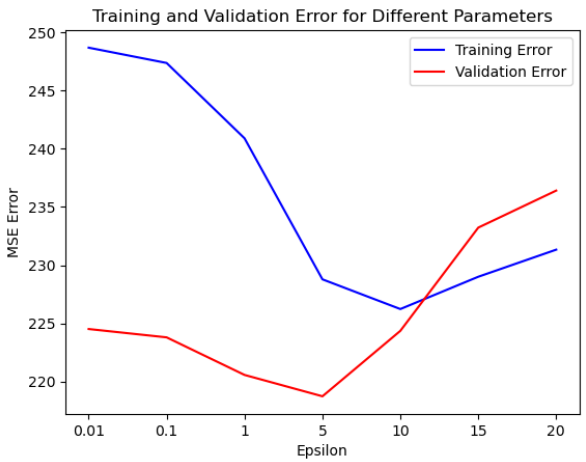
Support vector regression - MSE error



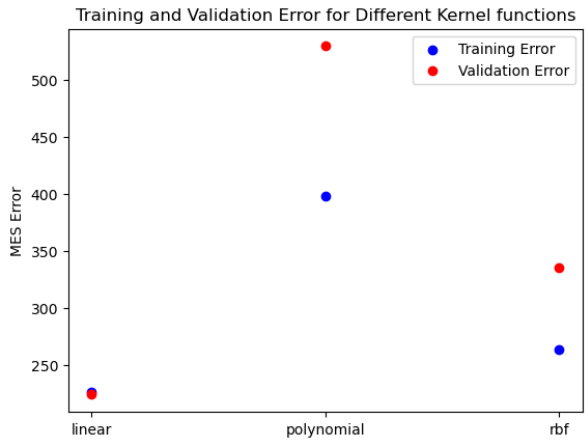
(a) Regularization



(b) Feature weights

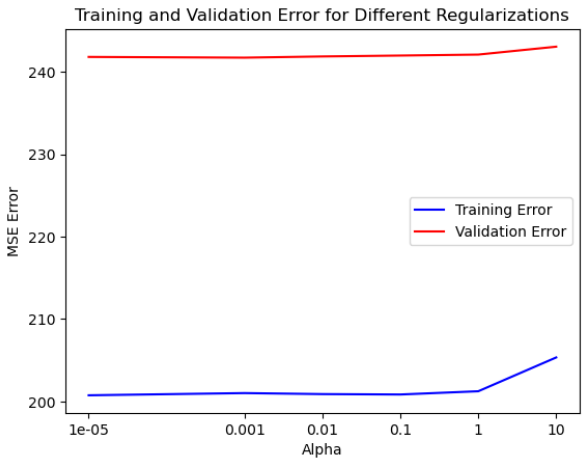


(c) Parameters

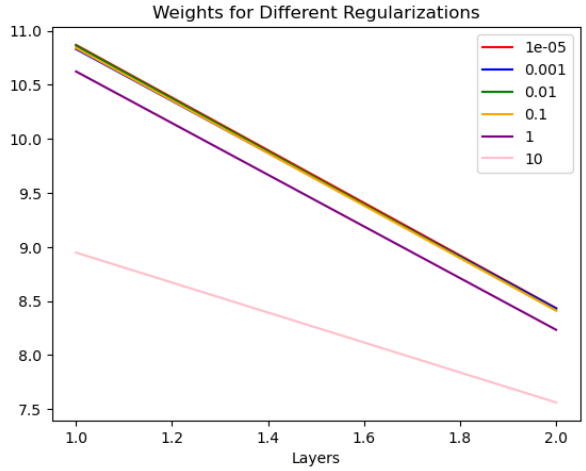


(d) Feature transformations

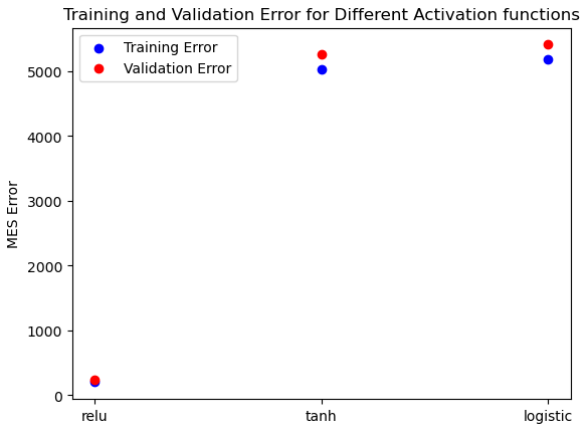
Neural networks - MSE error



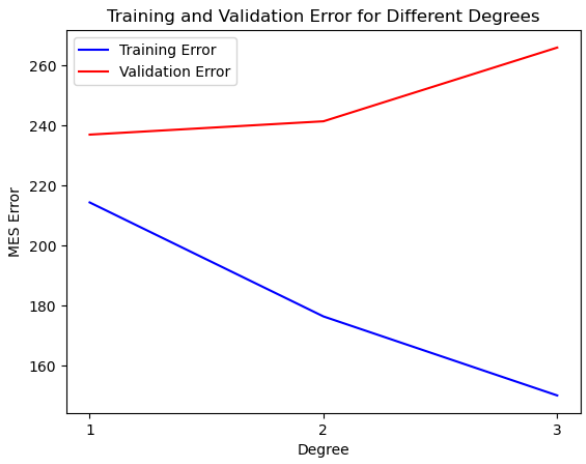
(a) Regularization



(b) Feature weights



(c) Feature transformations



(d) Polynomial degrees

Different normalization - MSE error

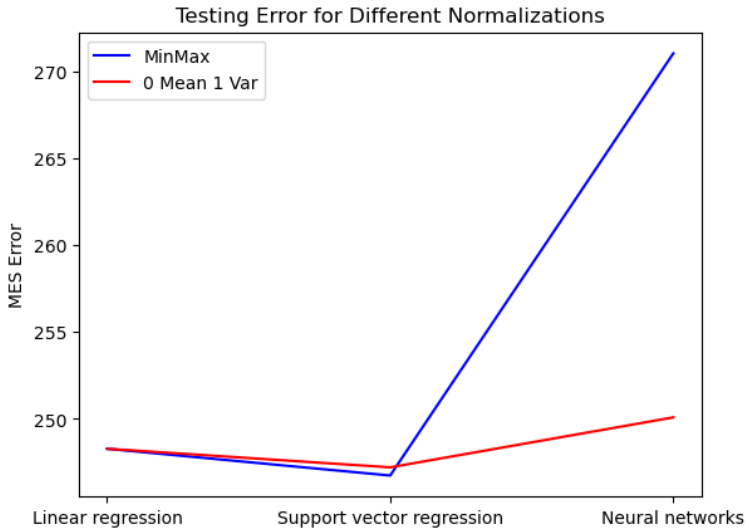


Table: Mean Squared Error for Test Set

Model	**MinMax**	**0 Mean 1 Var**
Linear regression	248.29	248.29
Support vector regression	246.75	247.22
Neural networks	271.02	250.09

Conclusion

Linear regression performed best without feature transformation.

SVR achieved optimal results with a linear kernel and moderate penalty ($C = 100, \epsilon = 5$).

Neural networks favored Relu activation and no feature transformation, $\alpha = 1$.

Overall, SVR with appropriate parameter seemed to be the most promising model for predicting airline ticket prices. 3 model's performance do not differ very much. Model is simple thus feature transformations is not necessary and Neural Network's performance is not very great. Most influncing feature is the price 1/2/3 days ago.