SEOULTECH
Seoul National University of Science & Technology

Assigment 2

20102122 정효안

Code	*ITM 526*	*Title*	*Business Analytics*		
Type	*Assignment*	*Questions*	*1*	*Weighting*	*5%*

1. Build prediction models using Rossman Store Sales Forecasting data

(1) Select input(explanatory) variables to predict sales. (10pts)

- List variables used in learning with explanation.

[Selected input variables: Promo, DayOfWeek, Assortment, StoreType]

- Promo: Promotions are likely to have a positive impact on sales
- DayOfWeek: I think the sales will be different depending on the day of the week
- Assortment, StoreType: Sales may be different depending on the store type and the assortment level. Because there are customers who prefer a specific store type or assortment level.

CompetitionDistance and Promo2 are also important for the following reasons.
But I will use new variables(CompetitionLevel,NowPromo2) made from these variables, not these variables.

- CompetitionDistance: The sales is likely to be inversely proportional according to the competition distance
- Promo2: Promotions are likely to have a positive impact on sales

(2) Add new variables (feature engineering) (20pts)

- Week of Year : I added it because it can reflect sales going up in a certain period of time each year (e.g. Christmas).
- NowPromo2: Whether the store is currently participating in promotion 2 or not
  (I added it because it was important to do promotion 2 at that time, not the present)
- NewPromo2: Whether a new round of promotions was started in the current month
- CompetitionLevel: The degree of competition according to distance. The higher it is, the more competitive
- CompetitionOpen: Months since Competition was open
- SalesPerCustomer: Average sales per customer

(3) Split data (10pts)

(4) Hyperparameter search with cross validation (only using training set) (40pts)

Ridge: best alpha=1.963828(At that time, R2 score is the highest and MSE is the lowest)
Lasso:best alpha=0.001(at that time,R2 is the highest, and mse and mae is the lowest)
(In lasso, I tried to run the code using logspace at once for two days like Ridge, but it didn't work because of the performance of the laptop, so I had to run it separately. Also, I tried to put the kfold code in the code to get r2, mse, and mae, but it was also impossible, so I saved the preprocessed data in the list(x_trains, x_vals) and ran it separately.)

(5) Performance comparison on the test set and draw conclusions. (20pts)

With the selected best hyperparameter, the performance of the test set was much improved compared to the performance during validation. The r2 score is higher, and both mse and mae are lower. As a result, we can see the importance of the decision of the hyperparameter by validation. This is because the performance of the model depends on the selected hyperparameter. (we can reduce the generalization error on unseen data)