Final Project Report - Aayam and David

For the final project, we decided to use the Rossman Stores Sales dataset. The Rossman Stores Competition involved predicting the future sales of 1115 Rossman Stores. The task of the competition was to create a model with their training dataset which included each individual stores' past sales and seventeen other predictor variables, and ultimately use their test dataset to predict what their future sales would look like for each individual store. We initially performed some EDA on the datasets to see how we could extract meaning out of the data, and more importantly, understand relationships between the predictor variables and the outcome variable itself.

Since the variable we were testing (Sales) was a continuous variable, we decided to adopt all the methods we had learned in this class for continuous variables and build several different models. We used four different models: Splines, Loess, Ridge Regression and LASSO. We hypothesized that Ridge Regression and LASSO model would be the most effective due to their shrinkage effect and were planning to use the score from the splines and loess as a baseline.

We finally ended up choosing Loess as the ultimate model. To do this our process was as follows:

- Since the train data was extremely large, which ultimately cause the runtime for our R-program to timeout even in rstudio, we decided to use a subset (10% at random) of the whole train dataset for the training each model.
- We trained the dataset on a Spline Model (see Extras section A). We chose competition distance as the predictor variable and tried to predict future sales. To find the best value of df to use, we performed a cross validation on the spline model to get the value of df that gave us the lowest RMSPE (root mean square percentage error). Using that df value, we trained the spline model and created the submission file and submitted it to kaggle.
- We took the same approach as the Spline model for Loess (check Extras section B). We performed CV to find the span value that gave us the lowest RMSPE. Using that span value, we trained the Loess model and got submissions and submitted them to kaggle.
- For Ridge Regression (check Extras section C), we not only used the competiton distance as predictor variable. We ended up using many different predictors (based on the EDA that we performed), since ridge regression inherently would shrink the beta-variables of the useless predictors. We performed CV, like the above models, to obtain lambda_star_ridge. Using that lambda value, we trained the ridge regression model, obtained a submission file, and submitted it to kaggle.
- We did the same as Ridge Regression for LASSO (check Extras section D).

Final Model:

We looked at the scores for all the models that we received on kaggle. Surprisingly, we received the best score for the Loess model with a span of 0.02, so we decided to choose that as the ultimate model for our final. Then we did the following:

- For our final model, we used the whole train data rather than using 10% of it.
- We performed a CV of loess model with span of 0.02 to see what score we would obtain on kaggle, which came out to be around 0.41.
- We trained our Loess model on the train data with span of 0.02.
- We used this model to made predictions on the test data and submitted it on kaggle.
- We got a score of 0.37!

We were really happy with our approach and results.