Group F: Abbas Shah, Kiryu Kawahata, and Caleb Brooks
Final Write-up

Introduction: Our project utilized the "Grocery Sales Forecasting" dataset, from Corporacion Favorita's Kaggle competition[1]. The problem in question was being able to predict consumer demand for retail products in brick-and-mortar grocery stores (the variable in question is called "Unit Sales", a continuous outcome variable representing the number sold of whatever unit a product comes in). [2] We obtained meta datasets from the Kaggle datasets that contained information about the stores, items sold, also used some outside data, reformatting a dataset available through the World Bank which contains a number of economic indicators at different times over the period covered in the Favorita dataset. The idea was to assess consumer purchasing power and ascertain whether that relates to number of units sold in the retail stores.

Models:
We started by making a baseline linear model that included a lot of predictors to get a rudimentary score we could compare our other, more sophisticated models against. We tried out stepwise regression and also intuited the usefulness of certain variables after our EDA. Then, we made regression trees (and made them more sophisticated through bagging and random forests), and found that while they were effective at displaying the relationship between Unit sales and different predictors (Figure 1), weren't better than our baseline linear model in terms of predictive power. Finally, we used LASSO – as, given the number of predictors available to us, the automatic variable selection provided by LASSO seemed a good choice. With the given training set containing five predictor variables, LASSO regularization (Figure 2) was able to produce a Kaggle score of 1.248. Interestingly testing the same method with more predictors in the metadata provides only a marginal improvement with a score of 1.171. Thus, LASSO was the model we ultimately ended up running with, as it had the best predictive power.

Some aspects of the metadata however have made evaluations a little challenging. Since there aren't enough complete observations with some metadata predictors for the test set, creating a prediction proves very difficult. It is also worth mentioning the presence of many missing "NA" values prevents the LASSO method from being tested. The only solution for this is either eliminating some predictors or all rows containing any missing values from the data, however ignoring entire observations for containing a missing value has a chance of increasing bias. After searching for many solutions it's apparent that LASSO is very difficult to perform on this dataset. Any modifications done to the NAs were rejected when attempting to predict results based on the model produced through LASSO.
We also briefly explored the possibility of using MARS (Multivariate Adaptive Regressive Splines), but due to the dataset lacking sufficiently explanatory continuous predictors for this technique to be effective, we decided against it.

Challenges: The first challenge we faced was in the sheer size of the dataset. The full set contained around 125 million observations, which proved too large to even be imported into R, let alone for us to make models. We ultimate worked with a pared-down version, with 1 million randomly sampled observation. Moreover, there was a lack of information about a lot of the variables in the Kaggle training dataset and the meta datasets, which made it hard to judge the kinds of variables available to us. It was

[1] https://www.kaggle.com/c/favorita-grocery-sales-forecasting

[2] The dataset is formatted as 1) one primary data table, with each observation representing the sales of a particular item at a particular location on a particular day, 2) a number of sets of metadata, containing information about item type, store location, holidays, and the price of oil. The contest is graded based on Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE) - normalized so that items which are normally bought in large quantities wouldn't skew the grading, and with perishable items given a larger weight than non-perishables.

also harder to use certain methods and models considering that there were a lot of categorical variables and relatively few numerical ones. The most significant problem with the information in the dataset pertained to the "item_nbr" variable, which identified individual products. It makes sense that this would be the most significant variable available to us, given that the nature of different products would lead to them being sold in very different amounts. Unfortunately, not all of the items contained in the test set are represented in the training set. We tried overcoming this in a few ways - namely, attempting to model without regard for the specific item, attempting to predict item type as an intermediary step, and partitioning the dataset so that we could generate predictions with "item known" and "item unknown" models. Finally, the dataset, given the nature of Unit Sales, Oil Prices, and other Economic Indicators, requires time series analysis to obtain more meaningful results.

Conclusion: Overall, the LASSO model was our best performing model, performing significantly better than the baseline linear model and the Regression Trees and Random Forests models. In the future, we would want to consider time series models to account for the impacts of time.

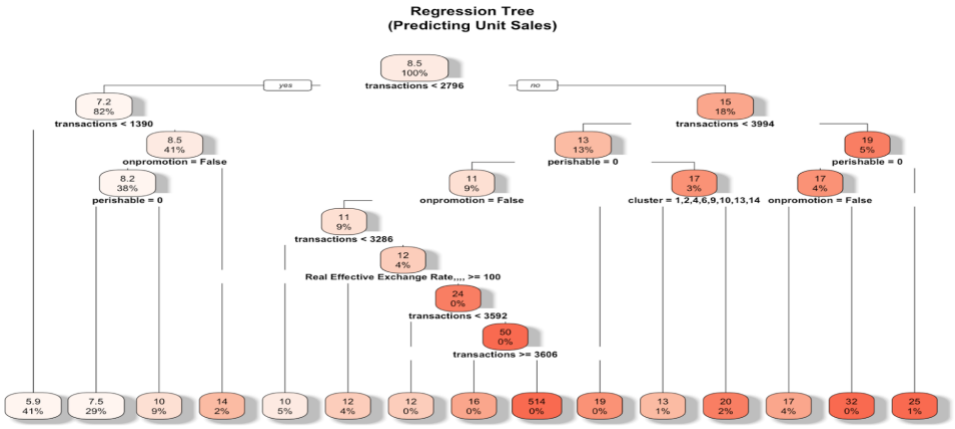

Figure 1: Regression tree to predict unit sales

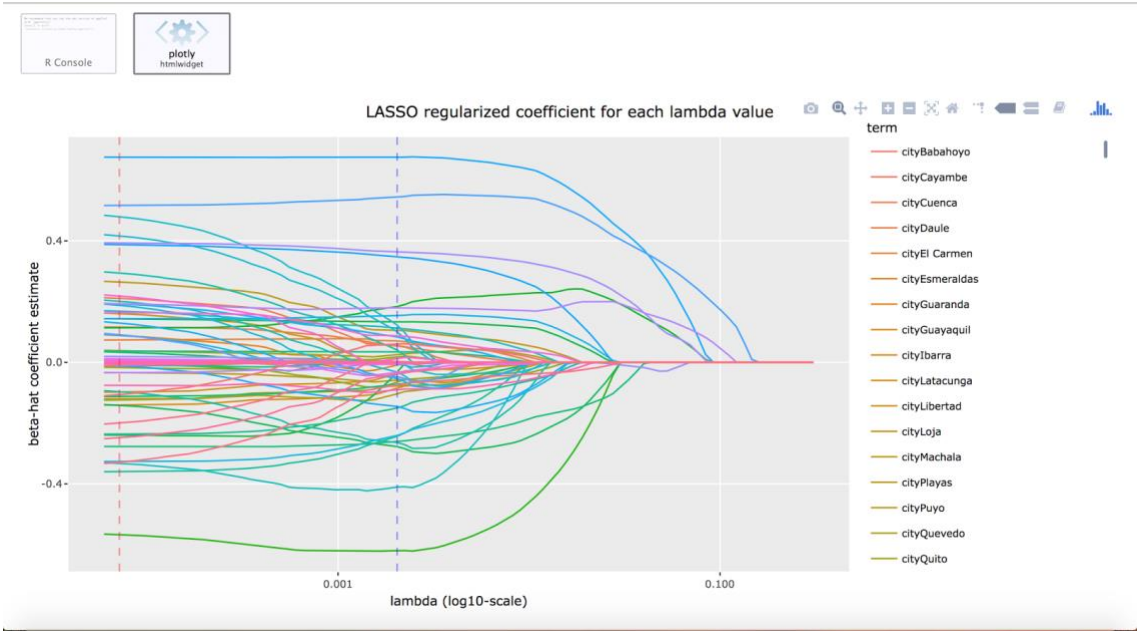

Figure 2: LASSO plot, estimated beta values plotted against log(lambda)