Hello Everyone my name is Yang Seunghyuck, today’s presenter

topic of 4th Assignment is Forecasting Sales, whish is the thing I failed. Let’s see together 😊

First, Explanatory variables.

The very first thing I did was drawing a correlation graph so I can get a general idea about whole factors at once. Through correlation graph, I noticed that Customer, Open and Promotion can be classified as first tier because their correlation rate is noticeably higher than other variables. And about other factors, I simply categorize them as second tier, which means I’m going to check them briefly after first tier.

Customer is the variable that has the highest correlation rate around here. The scatter plot and regression line are also well placed, apparently. So, I select it as variable. And about the case Open variable, first, it is binary data, which also has a clear result, absolutely tilted on one side. In addition , reasonable correlation number. But there was one thing I had to consider, which is about outliners. For deciding whether getting rid of them or not, I tried to count their number and found number of outliners are more than 200000, which is slightly more than 20% in total. Because it is quite huge amount, I judged it will effect on result if I discard them all. So, I had a decision not to transform, but just use them in vanilla state.

Next up is Promotion which is the last factor of tier 1. Binary data, Prominent correlation rate, and observable gap. But I discard promotion. The reason was standard deviation. As you can notice, standard deviation is too high even comparing with mean or other factors, I confirmed that this figure is abnormal. So I concluded Promotion would not be effective for forecast. Now let’s talk about tier 2 variables.

Day of week. obviously categorical data because one week is 7 days, But low correlation with Sales. And barely see any big difference in overall except for one day. Sunday, which is represented as number 7 in here. So I decide to transform this categorical data into dummy with 7 as standard.

And about other two things, State holiday and School holiday, I didn’t consider for choosing them. Because including graph, chart or statistic numbers, I couldn’t think it is related with Sales at all.

So this was my procedure to choose variables, and next I simply prepare data sets, including transforming and splitting into criterion, which is the year 2015. And here I have something to tell you. I was trying to evaluate factors through cross validation and spending time more than couple of days, but I failed. There were some problems related with my computer hardware for instance out of memory so I tried many things to get over, but I failed anyway. So, I unwillingly skipped this stage.

Last part. It’s about training model, compare for each. I choose Linear regression and KNeighbors regression. In here, also, I was choosing linear regression and Logistic regression. But Because of errors including memory problem, my computer is not capable for holding logistic regression. So, in explaining and summarizing at the same time, As you can observe from the graph and R^2 value, both trained models are quite accurate with predicting which is highly close to actual value. But still, maybe due to their features, the level of scattering showed a visible difference. Prediction from linear regression tends to spread out bit more, while one from KNeighbors narrow down.