In this homework, we were provided with 4 different data sets, one of which was a sample submission. We were asked to forecast the unknown sales data of the “Test set”, according to the datas given on “Train Set” and supplementary information on stores.

In order to predict those sales value, we have started by doing data preparing, we firstly added dummy variables for categorical variables on store data set, such as “PromoInterval”, “StoreType”, “Assortment” and added another variable that indicated if there exists a competition. After that we did data preparing on train data set by adding dummy variables to features such as “StateHoliday", “SchoolHoliday”. Then we merged those 2 datas together by inner merge on “Store” feature. After that we checked for any other improvements we could do, then we converted datetime data to 3 different features as “Day”, “Month” and “Year”. After that we thought we were ready so checked for multicollinearity and realized we have been fallen into the dummy variable trap, which might happen after adding dummy variables, so we fixed that by dropping a column.

After this we thought we were ready for regression and ANN but we have talked and realized that even though our values are numerical, it does not mean they are not categorical. Features like “DayofWeek”, “Day”, “Month”, “Year”, “Promo2SinceYear”, “Promo2SinceWeek”, “CompetitionOpenSinceMonth”, “CompetitionOpenSinceYear” and “Store” were actually categorical, since they are not representing any amount, they are not quantitative. So again started adding dummy variables for those values. We did the process we did earlier and handled this situation. We forgot to check for multicollinearity or check the test set for its given columns and went straight on with the regression and ANN model fitting. After those steps we fitted and prepared an ANN model. After that we found an MSE value of 158.706,18 and an MAE value of 263,15 on our validation data which we splitted with using sklearn’s train\_test\_split class. Until this point, these values seemed good so we were preparing the test set for prediction, then we realized that the test data set had no column as “Customers” on the given data. If we added an empty column with values of 0 for customers, we would get a very inaccurate assumption since “Customers” feature is one of the most important features for our model. So we went back and dropped the “Customers” column from the data set and fitted the model again.

This time we found an MSE value of 1.835.970,34and an MAE value of 989,74 on our regression model, and an MSE value of 1.408.200,5 and an MAE value of 783,64. These values were not looking good so we went back and started talking what could we do. We realized we haven’t checked for multicollinearity after adding all those dummy variables the second time so we went back and checked for multicollinearity again. Since we added tons of columns (around 150 more columns), this checking process in python was taking a long time to run and we do not recommend running it on your computer since it may take around 5 minutes. After a while, the output was shocking since more than half of our variables was collinear. We tried to fix it by dropping some features but we could only fix some of them. After dropping them we went on with the code again, fitting models, predicting and evaluting errors, we got an MSE value of 1.835.941,94 and an MAE value of 989,72 for our regression model and an MSE value of 1.471.432,25 and an MAE value of 804,40 on our ANN model.

After looking at these numbers we can say that, even after trying to fix multicollinearity, we still could not manage to decrease our error. So we ignored the fixing multicollinearity part and went with our second model (since our first model included the column of “Customers” on it and we did not have that column on our test data set which we are going to predict the Sales values of). After prediction, we have exported our values to a .xlsx file like the sample submission. This file is included on our uploaded files.

As we found the mean value for our predicted sales as 1489, we think we did not did a very good job on forecasting these test values. It might be becuase of those extra columns in our training set such as extra store types. If we did this from scratch over again, we would check for the test set columns first and include only those important columns in our model, instead of including every single one on the training set.

070180708 Salim Kaplan

070180114 Batıkan Soysal

070180133 Kerem Karagöz

070180147 Yakup Emre Gülhan