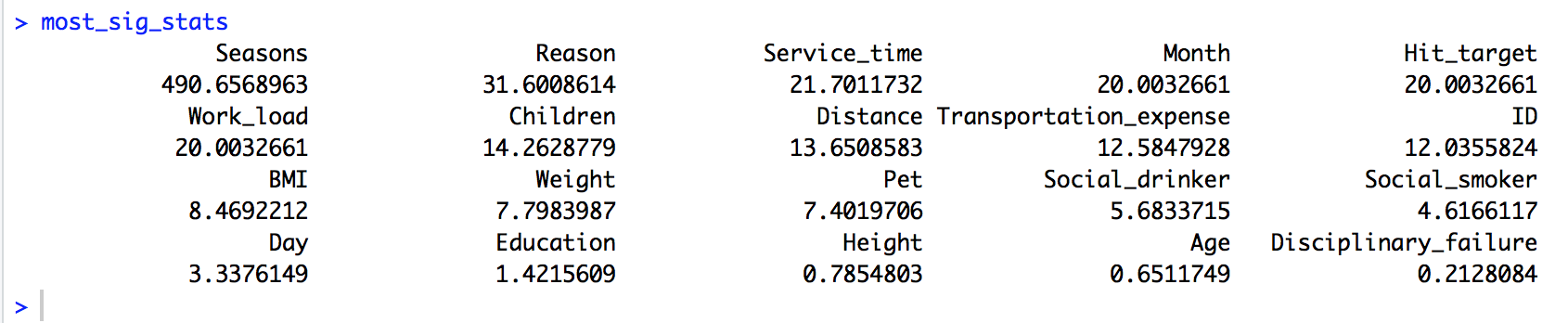
Refining of KNN model conclusions.

In order to refine our model, one of the techniques we used, involved cross validation in order to select the optimal number of variables and k. Let’s recap, our database, consisted in a matrix of 740 rows by 21 columns. From those 21 columns, predictors were actually 19 since one of them is the response variable; ‘Absent\_time’ and one column is the ID of each employee.

***Analysis of the relevant variables:***

The first part of the refining process, consisted in the selection of the variables, since we have 19 predictors, some of them might not be relevant in the classification process. This goal was achieved by fitting linear model looping over each variable, and performing F-statistic tests. The results of the test are shown below, with the top 5 variables being: ‘Seasons’, ‘Reason’, ‘Service\_time’, ‘Month’ and ‘Hit\_target’.



***Finding optimal number of variables and k.***

Determining the optimal number of variables to use and k, we employed Monte Carlo cross validation technique. We used replication loops with the objective to compute both mean errors and mean sensitivity for all combinations of number of variables and k. We analyzed two metrics: error and sensitivity however, we believe that, in this case sensitivity is the most important since we are trying to predict accurately the positive class and we just have a small number of them. The graphs below, show all the results for mean errors and mean sensitivity for all the combinations of K and number of variables. This means that, values of k between 1 and 3 and variables between 5 and 15 seem to generate the best results however, in order to keep the model as simple as possible, we believe that k=3 and number of variables 5 would be the optimal model to proceed with.

