**KNN**

K- nearest neighbour (kNN) is used for classification and regression which uses non-parametric method. After exploring the clickstream dataset, we used kNN classification method in traing the dataset. It uses the equation below to measure the Euclidean Distance and classify the data:

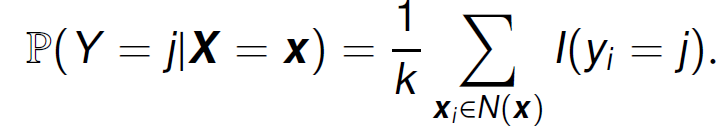


Figure 2.1: Formula for KNN

K value is crucial in the kNN model. A very small K value will allow noises to influence the final result, whereas a very high K value will result in underfitting the model. Therefore, we used the square-root method and the k-fold cross validation method to select the optimal K value.



Figure 2.2: Too many ties error using kNN

Since the dataset is very large (165474 rows at 14 variables), the normal kNN model using the R library “class” is unsuitable for training the dataset because it results in the error “too many ties”. Hence, we used the library “kknn” and performed weighted kNN which give more weight to nearby points and less weight to the further points in order to train the dataset.

The square root method in selecting the K value takes the value of the number of rows of the stratified sampled training data (in this case, 75% training data was used) and square rooting the value. This method gives us a K value of 353.

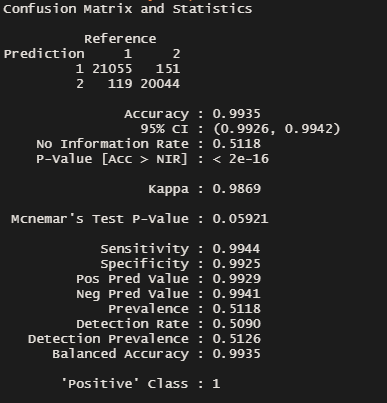


Figure 2.3: Confusion Matrix of Weighted KNN using kknn

Figure 2.3 shows the confusion matrix of the weighted kNN model with K = 353, it gives us an accuracy of 99.35%.

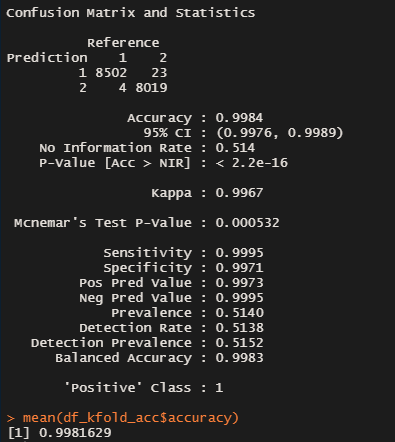


Figure 2.4 Confusion Matrix for One of the 10 folds with Weighted KNN Model

Figure 2.4 shows one of the confusion matrix of the weighted kNN model with K = 353 which uses the k-fold validation method. We applied a 10-fold validation method and this gives us an overall accuracy of 99.81629%. In conclusion, the 10-fold validation method performs better than the stratified sampling using square root method in the kNN model.