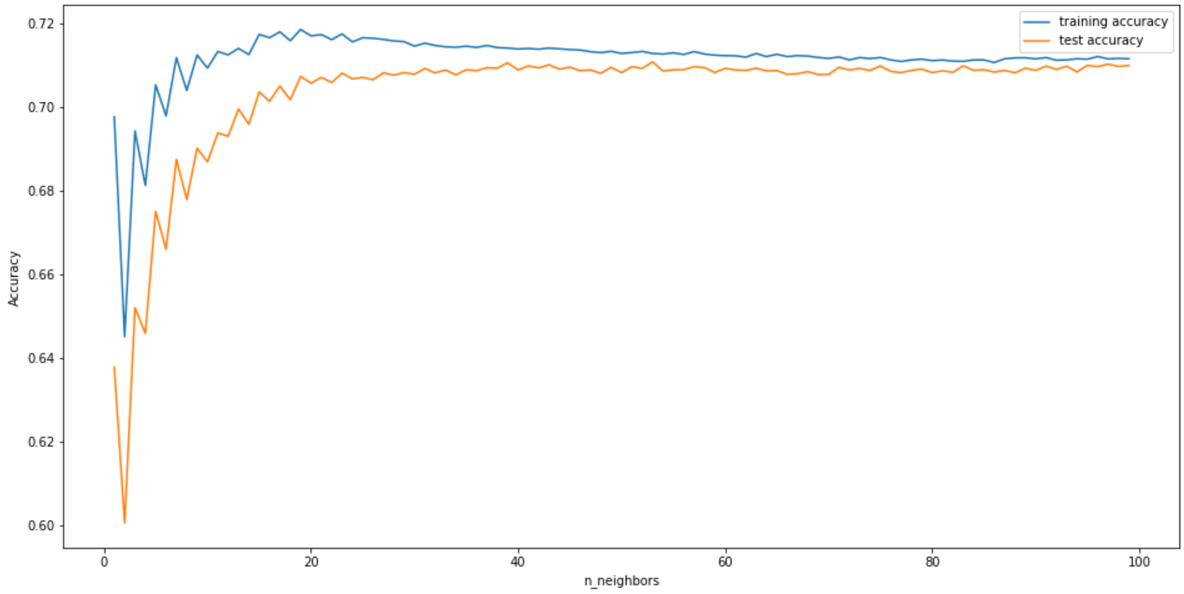
The final model we used to predict the Boolean that the project was funded is kNN model. The features used to train the model are normalized USD_goal, staff_pick and one-hot encoded categories. The following graph shows the validation curve about accuracies of both training dataset and test dataset with the split ratio 7 : 3.

From the graph, we can observe that the training and test accuracy curve converged at about k equals 100 and the test accuracy is around 0.70.



According to the figure above, we can say that the model has very low variance but a moderately high bias if the success criteria is defined as an accuracy above 0.8. But to us, the result is fairly acceptable.

We also conducted cross-validation tests with 200 trials and the mean of the validation score is roughly 0.695 which is very close to our randomly split dataset test result.

Based on the fact that the normal distribution of log-transformed USD_goal in the test dataset is very similar to it in the training dataset and this feature is considered to be the leading feature that influences the predicted result, we have confidence that the predicted accuracy should be close to the cross-validation result. Therefore, hinged on the above reasons and plus unexpected bias, we think the accuracy of the prediction is about 0.69 and this leads to the number of correct predictions = 78065*0.69 ≈ 53865.