**Consider the models**

In this assignment I created and compared the following models on the scaled as well as unscaled dataset provided:

1).Logistic Regression: 2)Knn: 3)Random Tree classifier. Also tried SVM but it was computationally expensive and hence got rid of the SVM model.

Started with an assumption that the Logistic will work better in dataset provided.

Started with investigating and preprocessing of dataset. The dataset was well balanced. Categorical dataset was converted to numeric and then target and feature were separated.

UnScaled dataset

Comparison between logistic/KNN/randomforest

1)**logistic** 0.6508210180623973 0.5168013611229264

2)**Knn** 0.6463054187192119 0.5110591237771162

3)**random forest** 0.9994252873563219 0.6371756699276904

4)**random forest with feature selection0**.9995894909688013 0.5935772011909825)**Logistic With feature selection from Random Forest** 0.6490968801313629 0.5180774138664398

The values of training/testing score indicate that logistic is the best model for out data set.

**Revisit the Preprocessing: Scale the data**

The data going into these models was never scaled, an important step in preprocessing. Use StandardScaler to scale the training and testing sets.

Scaling of the dataset should get us a better model as it will prevent model to be biased by varied magnitude values of different features

Fit and score the LogisticRegression ,Knn and RandomForestClassifier models on the scaled data.

Comparison between logistic/KNN/randomforest/randomforest selected features/logistic on selected features

**1)logistic** 0.7078817733990148 0.767333049766057

2)KNN 0.683743842364532 0.5572096980008507

3)random forest 0.969047619047619 0.6552530837941302

4)**random forest selected features**  0.9945812807881773 0.5765631646108039

5)**Logistic Selected features**  0.7076354679802955 0.78285835814547

We can see here too the logistic works best for the data set provided. Scaling the data improves the model accuracy and gives better model. I tried random forest with/ without selected features but it the accuracy scores were not good. I thought using logistic model with selected feature can help improve the accuracy but the results(scores) indicated otherwise. Probably could be that this a well balanced data and also a less noisy data set