Credit Card Default Fraud Prediction
Artificial Neural Network and K Nearest Neighbors

Introduction

In this project, we will be implementing Artificial Neural Network and K Nearest Neighbors Credit card payment default dataset. We will use k fold cross validationand parameter tuning to identify the best algorithm.

The dataset source of Credit card default dataset:
https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.It has 30,000 observations and 25 attributes. The target variable is whether credit card user will default or not.

Credit Card Default Dataset

Artificial Neural Network (ANN)

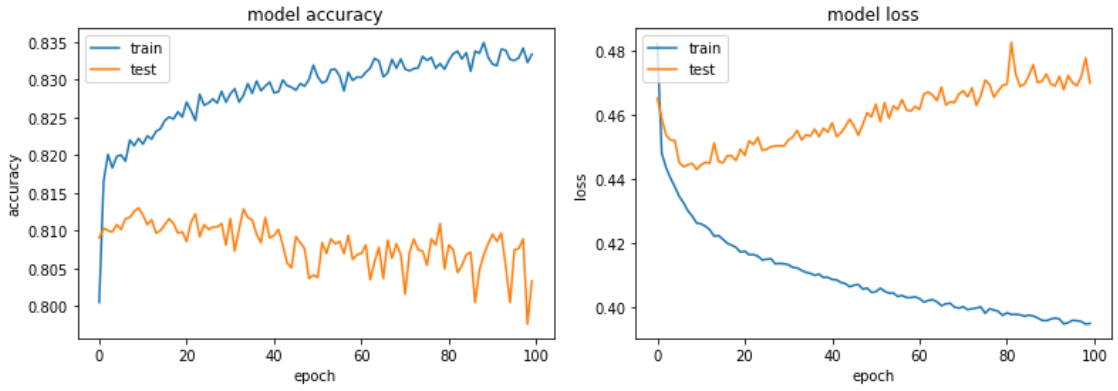
Running Artificial Neural Network created using Keras by TensorFlow, using Sigmoid as the output function for binary classification.

Experimenting with number of epochs

Epochs is the number of times the dataset passes through the neural network to update beta coefficients and weights of the nodes. Created ANN with 2 layers of 16 and 32 nodes and 'relu' as activation function for each layer. Ran the neural network for 3 values of epochs which are 100,250 and 500.

Epochs	Test Accuracy	Precision score
100	0.81	0.35
250	0.80	0.34
500	0.81	0.33

The best accuracy and best precision score are obtained after 100 epochs. We will be using number of epochs as 100 for subsequent experiments.

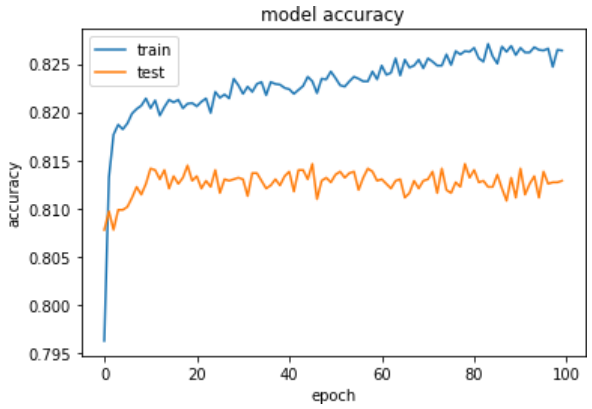


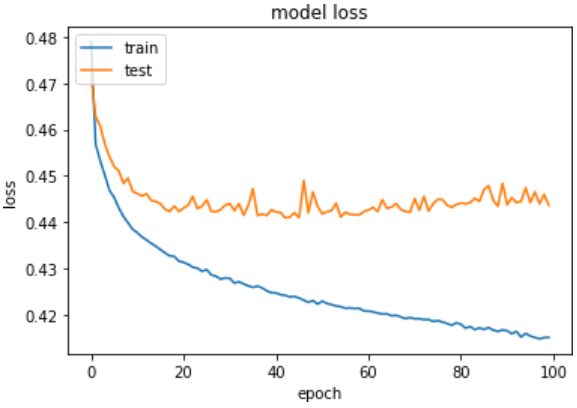
From model accuracy plot we can see model accuracy increases as number of epochs increases. The test accuracy decreases, but not significantly. The test accuracy seems to have low variance indicating model is generalizing well. From model loss plot we can see that the training loss reduces as number

of epochs increases. Testing increases by only slight amounts as number of epochs increases. This indicates themodel has low bias.

Experimented with number of layers and different activation functions. Best model was obtained from ANN with 2 layers. 'SoftMax' activation function is used for input layer and hidden layer. 'Sigmoid' function for the output layer.

The best accuracy obtained is 82% with precision score of 0.37. Model accuracy plot and model loss plot are shown below:



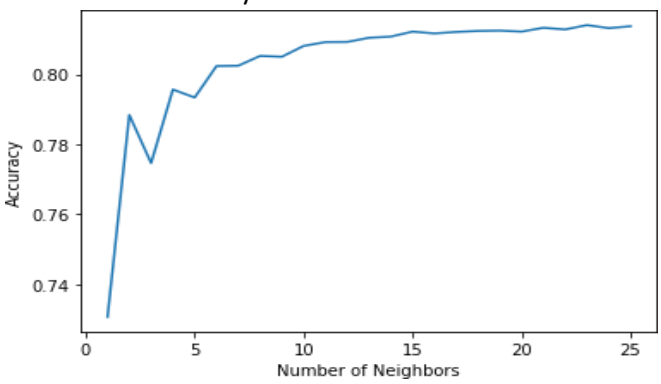


We can see from the above plots that model test accuracy and model test loss are almost constant. Indicating low variance and low bias. Hence this is a good model.

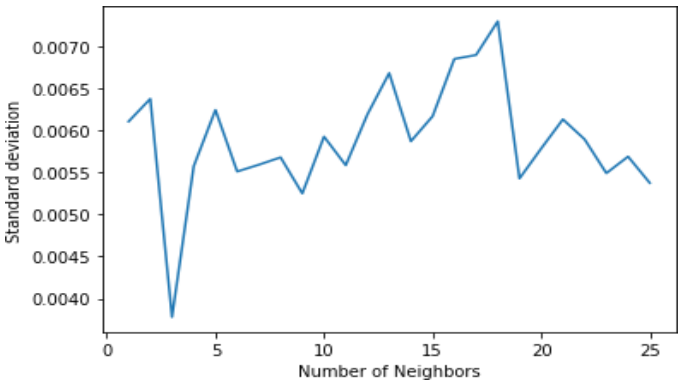
K Nearest Neighbors (KNN)

Selecting the best number of neighbors k for KNN model using K fold cross validation. Experimented with 25 values of k.

Plot of accuracy vs k



Plot of standard deviation vs k

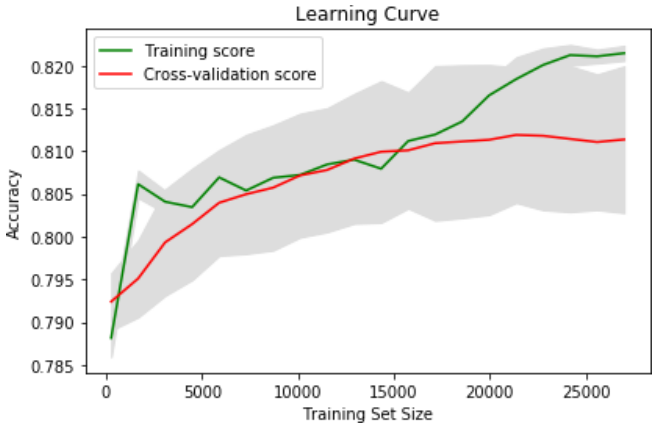


k=23 gives the best accuracy of 81.2% and lowest standard deviation of accuracies from K fold CV. The model has a precision score of 0.30

Experimenting with different distance metrics

Distance metric	Test Accuracy %	Precision score
Manhattan	81.5	0.30
Chebyshev	80	0.30
Euclidean	81.06	0.30

The model with best accuracy is the one with Manhattan distance metrics and number of nearest neighbors being 23.



From the learning curve we see that as training size increases the gap between validation accuracy and training accuracy increases indicating the model has high bias.

Model Comparisons

Model	Test Accuracy	Precision Score
SVM	82%	0.34
Decision Tree	73%	0.34
Boosting	82%	0.34
ANN	82%	0.37
KNN	81.5%	0.30

The best model for the Credit Card Default dataset would be ANN model with test accuracy of 82% and highest precision score of 0.37.