Luddy School of Informatics, Computing and Engineering

Home Credit Default Risk

Team Members:

1. Akash Gangadharan (akganga@iu.edu
2. Laya Harwin (lharwin@iu.edu)
3. Dhairya Shah (shahds@iu.edu)
4. Shobhit Sinha(shosinha@iu.edu)

Agenda For Phase 4

01 Project Timeline

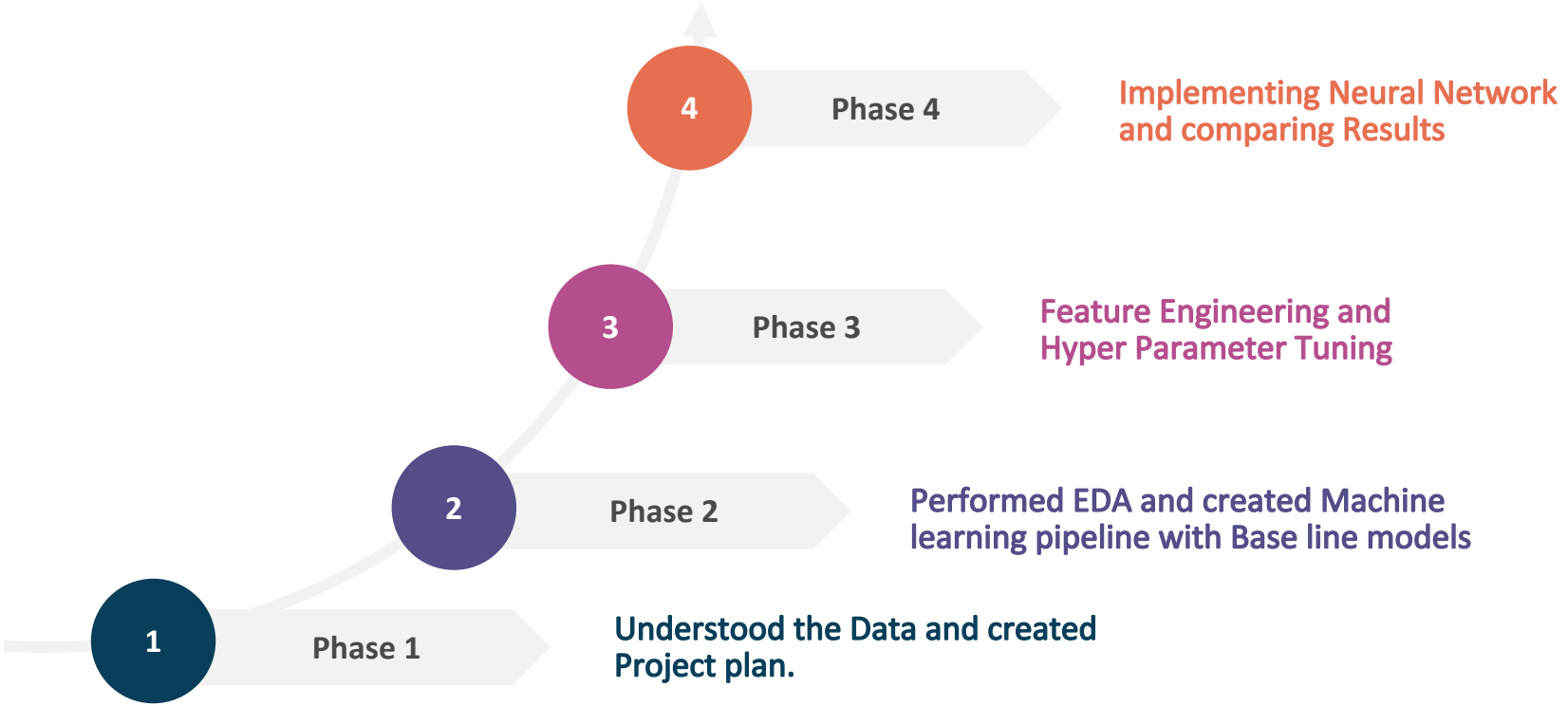
02 Feature Engineering and Hyper Parameter Tuning

03 ML Project Map

04 Comparing results and Conclusion

05 Kaggle submission

Project Timeline



Feature Engineering

Treating Highest proportion of zero values and Dividing main data into categorical and numerical features

Treating missing values and correlation with respect to target variables

Adding 17 new features

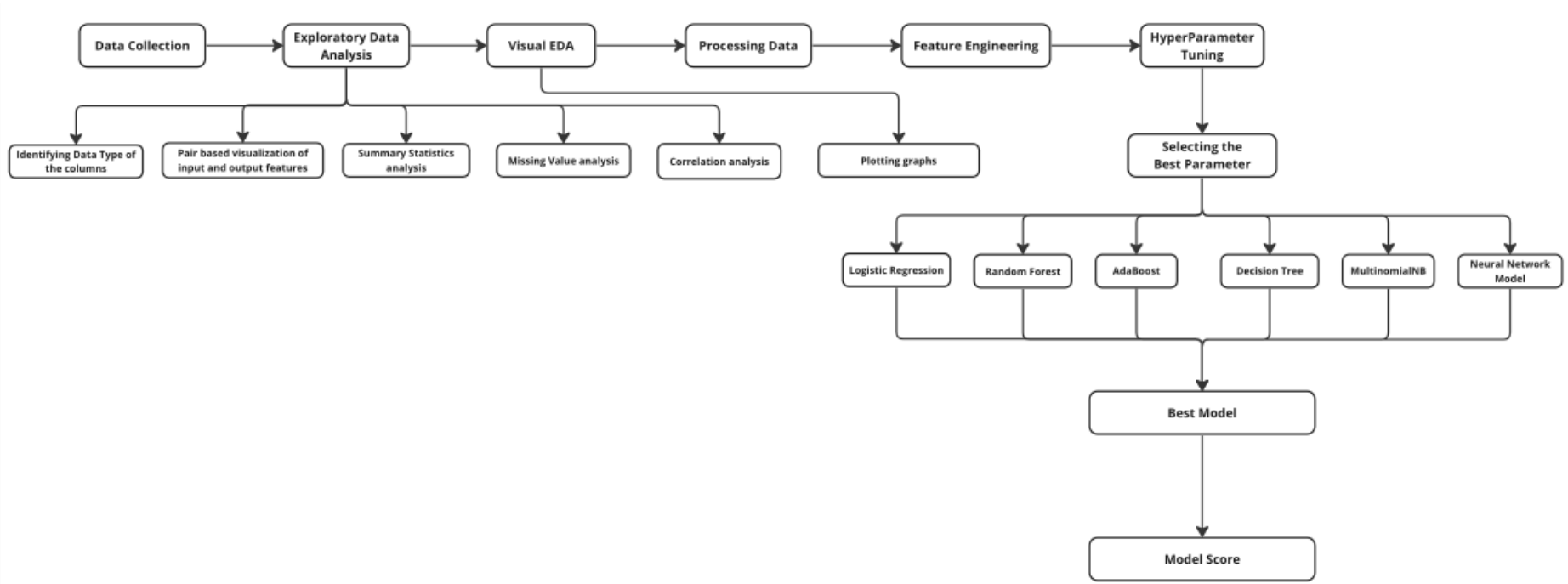
Performed One Hot Encoding.

Hyper Parameter Tuning

Grid Search CV was used to find the best parameters on the following models

- Logistic Regression
- Decision Tree
- Random Forest
- AdaBoost
- Neural Network

ML Project Map



Experiment 1:Single Layer Perceptron

Architecture	No of Input Features	Families of Input Features and Count per family	Loss Function	Hyper Parameter Setting	Train CXE Loss	Train Accuracy	Train ROC AUC Score	Valid CXE Loss	Valid Accuracy	Valid ROC AUC Score	Test CXE Loss	Test Accuracy
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 500	365.503784	0.726961	0.519934	362.846497	0.728415	0.521210	368.333618	0.725756
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 600	875.312744	0.690097	0.518087	867.004700	0.691423	0.519682	880.751770	0.688862
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 700	13.264116	0.849517	0.527121	13.192088	0.849045	0.527181	13.656445	0.846878
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 1500	284.093414	0.729131	0.510399	283.088989	0.727561	0.506515	285.184906	0.726732
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	153.057281	0.907739	0.498929	152.066010	0.907480	0.499088	285.184906	0.726732
Linear(in_features=17, out_features=2, bias=T	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 600	322.381042	0.907805	0.498544	321.221436	0.907561	0.498902	330.176758	0.907041

Experiment 2:Multi Layer Perceptron Classification

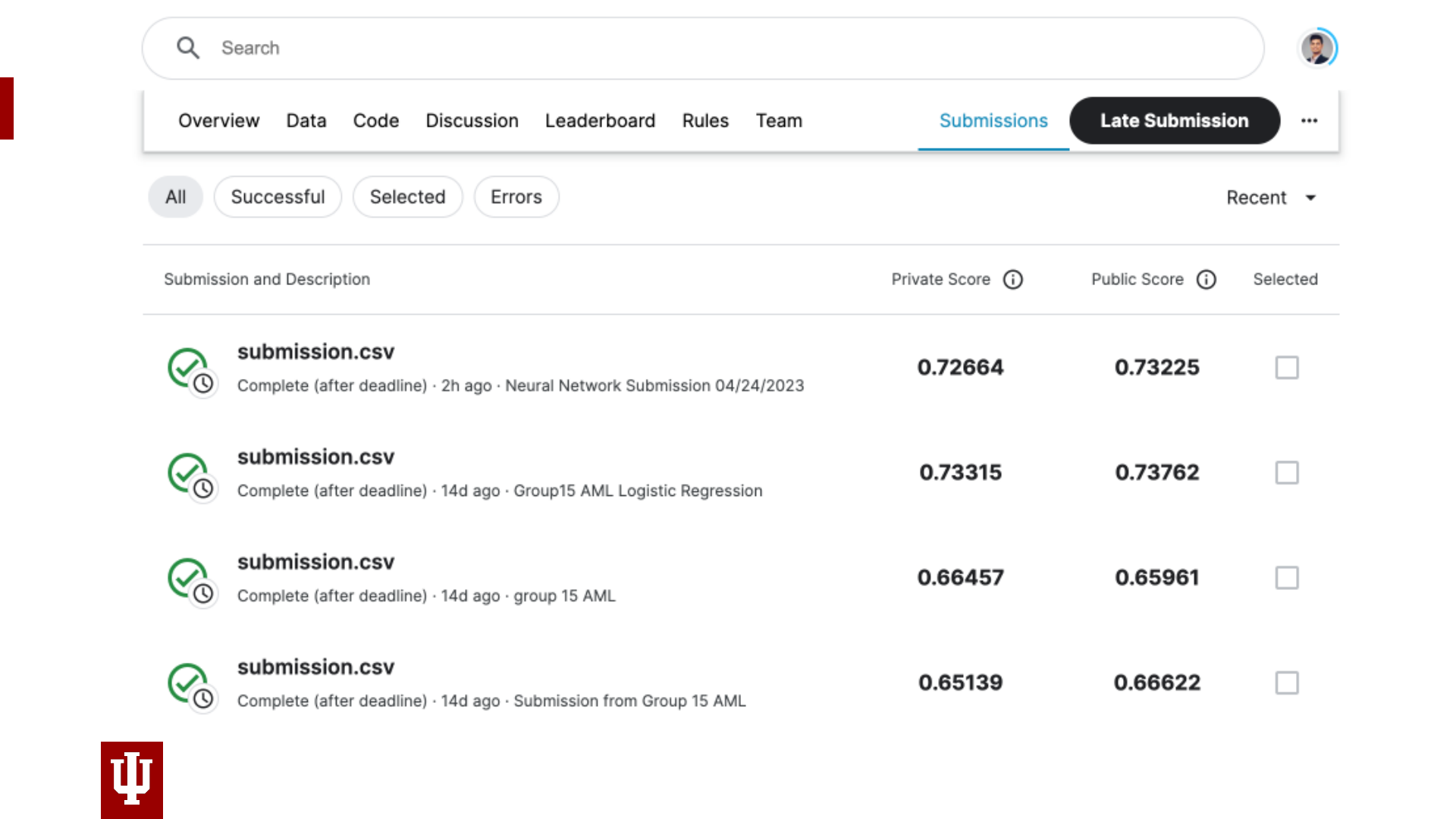
Number of Hidden Layers	Number of Neurons in each layer	No of Input Features	Families of Input Features and Count per family	Loss Function	Hyper Parameter Setting	Train CXE Loss	Train Accuracy	Train ROC AUC Score	Valid CXE Loss	Valid Accuracy	
[Linear(in_features=17, out_features=50, bias=	5 [50,40,30,20,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 500	3.127162	0.331306	0.477238	0.923185	0.918028	0
[Linear(in_features=17, out_features=50, bias=	5 [50,40,30,20,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 700	0.352160	0.919558	0.500000	0.352199	0.919390	0
[Linear(in_features=17, out_features=100, bias	7 [100,90,80,60,50,30,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.01, epochs: 700	0.279848	0.919558	0.500000	0.280256	0.919390	0
[Linear(in_features=17, out_features=100, bias	8 [100,90,80,60,50,30,10,5]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.01, epochs: 800	0.279848	0.919558	0.500000	0.280256	0.919390	0
[Linear(in_features=17, out_features=80, bias=	6 [80,70,60,50,40,30]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.279566	0.919558	0.610168	0.285471	0.919390	0
[Linear(in_features=17, out_features=40, bias=	7 [40, 35, 27, 20, 15, 10, 5]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.276272	0.919558	0.609742	0.276299	0.919390	0
[Linear(in_features=17, out_features=200, bias	10 [200,180,150,125,100,80,75,50,25,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.277960	0.919558	0.614260	0.282230	0.919390	0

Experiment 3: Multi Layer Perceptron Regression

Number of Hidden Layers	Number of Neurons in each layer	No of Input Features	Families of Input Features and Count per family	Loss Function	Hyper Parameter Setting	Train CXE Loss	Train Accuracy	Train ROC AUC Score	Valid CXE Loss	Valid Accuracy
[Linear(in_features=17, out_features=50, bias=	5 [50,40,30,20,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.0001, epochs: 500	3.127162	0.331306	0.477238	0.923185	0.918028
[Linear(in_features=17, out_features=50, bias=	5 [50,40,30,20,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 700	0.352160	0.919558	0.500000	0.352199	0.919390
[Linear(in_features=17, out_features=100, bias	7 [100,90,80,60,50,30,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.01, epochs: 700	0.279848	0.919558	0.500000	0.280256	0.919390
[Linear(in_features=17, out_features=100, bias	8 [100,90,80,60,50,30,10,5]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.01, epochs: 800	0.279848	0.919558	0.500000	0.280256	0.919390
[Linear(in_features=17, out_features=80, bias=	6 [80,70,60,50,40,30]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.279566	0.919558	0.610168	0.285471	0.919390
[Linear(in_features=17, out_features=40, bias=	7 [40, 35, 27, 20, 15, 10, 5]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.276272	0.919558	0.609742	0.276299	0.919390
[Linear(in_features=17, out_features=200, bias	10 [200,180,150,125,100,80,75,50,25,10]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	lr=0.001, epochs: 500	0.277960	0.919558	0.614260	0.282230	0.919390

Experiment 4: Multihead model CXE and MSE

Number of Fully Connected Layers	Number of Neurons in each layer	No of Input Features	Families of Input Features and Count per family	Loss Function y1	Loss Function y2	Hyper Parameter Setting	Train Y1 Accuracy	Valid Y1 Accuracy	Test Y1 Accuracy	Train Y1 ROC AUC	Valid Y1 ROC AUC
MultiHeadModel(\n (fc1): Linear(in_features=1	2	[16,8]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 150, Optimizer	0.919312	0.920163	0.918276	0.500000
MultiHeadModel(\n (fc1): Linear(in_features=1	3	[20,10,8]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 300, Optimizer	0.080688	0.079837	0.081724	0.500000
MultiHeadModel(\n (fc1): Linear(in_features=1	3	[32,16,8]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 301, Optimizer	0.288509	0.285512	0.290309	0.480564
MultiHeadModel(\n (fc1): Linear(in_features=1	6	[128,64,32,16,8,4]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 701, Optimizer	0.080688	0.079837	0.081724	0.500000
MultiHeadModel(\n (fc1): Linear(in_features=1	7	[128,64,32,16,8,4,2]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 301, Optimizer	0.080688	0.079837	0.081724	0.500000
MultiHeadModel(\n (fc1): Linear(in_features=1	9	[512,256,128,64,32,16,8,4,2]	17	Numerical Features: 8, Categorical Features: 9	CrossEntropyLoss	MSE	Learning rate = 0.001, epochs = 301, Optimizer	0.080688	0.079837	0.081724	0.500000


Search
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
Submissions
Late Submission
All
Successful
Selected
Errors
Recent
Submission and Description
Private Score
Public Score
Selected
submission.csv
Complete (after deadline) · 2h ago · Neural Network Submission 04/24/2023
0.72664
0.73225
submission.csv
Complete (after deadline) · 14d ago · Group15 AML Logistic Regression
0.73315
0.73762
submission.csv
Complete (after deadline) · 14d ago · group 15 AML
0.66457
0.65961
submission.csv
Complete (after deadline) · 14d ago · Submission from Group 15 AML
0.65139
0.66622

Key Results/Findings

In phase 4, we built a NN pipeline and experimented with different architectures while avoiding leakage and cardinal sins. We used PyTorch and adjusted training parameters and detected leakage by comparing categorical features of training and test datasets.

Comparing all the results that we got from the experiments that we conducted we conclude that Logistic Regression achieved 91.85% accuracy with feature engineering only. The second highest test accuracy was of multi-layer perceptron (MLP) neural network: Classification that was 91.83%.

Thank You!

INDIANA UNIVERSITY BLOOMINGTON