Assigment – Rohit Garg

1. Objective

The objective of this exercise is to do binary classification on the data provided.

- Train data has 3,910 observations and 58 variables
- Test data has 691 observations and 57 variables

2. Data

The model is developed on 3,910 observations and 57 independent variables. There are 2,376 (61%) non-events and 1,534 (39%) events. It is observed that there is no missing values in the independent variables.

2.1. Shortlisting of variables

Based on information value (IV) 9 independent variables are shortlisted. IV indicates the explanatory power of the independent variables.

- IV = inf indicates that for a particular bin all the observations are either event or non-event
- 9 variables are – X27, X29, X25, X7, X26, X42, X24, X31 and X23

Variable	IV	VIF
X27	inf	1.115
X29	3.357	1.841
X25	3.090	1.563
X7	2.901	1.123
X26	2.812	1.516
X42	2.667	1.461
X24	2.653	1.088
X31	2.451	2.379
X23	2.188	1.322
X20	1.898	1.134
X53	1.815	1.233
X35	1.651	1.720
X30	1.578	2.184
X43	1.516	1.233
X56	1.505	2.398
X55	1.487	1.424
X15	1.382	1.357
X46	1.348	1.210
X16	1.283	1.069

Variable	IV	VIF
X57	1.251	1.635
X45	1.213	1.094
X28	1.210	1.940
X11	1.101	1.206
X52	1.064	1.082
X44	0.941	1.036
X8	0.829	1.104
X37	0.781	1.281
X21	0.775	1.327
X39	0.772	1.146
X17	0.770	1.200
X2	0.691	1.033
X9	0.674	1.233
X6	0.648	1.114
X50	0.646	1.637
X5	0.565	1.131
X54	0.539	1.078
X33	0.506	1.059
X19	0.460	1.320

Variable	IV	VIF
X36	0.452	2.656
X3	0.434	1.111
X10	0.434	1.087
X18	0.422	1.159
X1	0.416	1.115
X12	0.406	1.111
X13	0.390	1.096
X14	0.388	1.078
X51	0.234	1.087
X49	0.223	1.249
X40	0.020	3.534
X4	0.000	1.006
X22	0.000	1.300
X32	0.000	110.333
X34	0.000	107.190
X38	0.000	1.047
X41	0.000	1.172
X47	0.000	1.014
X48	0.000	1.015

2.2. Descriptive statistics

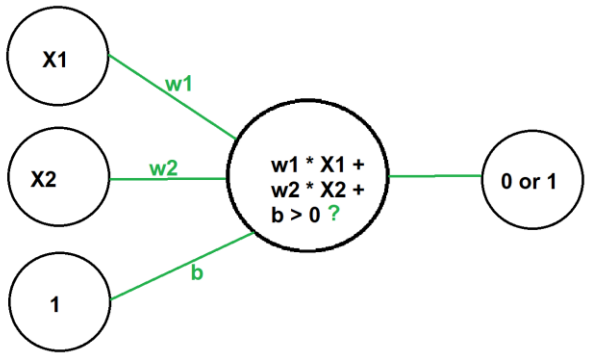
	count	min	mean	max	std
X27	3,910	0.000	0.757	33.330	3.322
X29	3,910	0.000	0.100	14.280	0.608
X25	3,910	0.000	0.566	20.830	1.734
X7	3,910	0.000	0.112	7.270	0.390
X26	3,910	0.000	0.267	16.660	0.893
X42	3,910	0.000	0.139	14.280	0.800
X24	3,910	0.000	0.095	12.500	0.443
X31	3,910	0.000	0.066	12.500	0.409
X23	3,910	0.000	0.101	5.450	0.346

3. Multi-layer Perceptron

3.1. Methodology

Deep learning is the science to allow computers to learn just like humans, particularly learn patterns from information. Machine learning has supervised, unsupervised and semi-supervised algorithms. Deep learning is a part of machine learning. There are specific algorithms that are a part of deep learning. Deep learning consists of a stack of layers consisting of neurons and activation function.

- **Supervised algorithm:** Teaching the algorithm using inputs and outputs. The output is the label identifying fraud and not fraud.
- **Feature extraction:** Extracting the most valuable features.



- **Objective:** To find the most optimal values for w1, w2 and b
- **Inputs:** X1 and X2
- **Weights:** w1 and w2
- **Bias:** b
- **Output:** 0 or 1

3.2. Hyper-parameters

GridSearchCV exhaustively considers all parameter combinations. It is used for tuning the hyper-parameters of an estimator. The GridSearchCV instance implements the usual estimator API, when "fitting" it on a dataset all the possible combinations of parameter values are evaluated and the best combination is retained.

Iteration	Set (CV=5)	Selection	AUROC
1	**hidden_layer_sizes:** (9),(9,9),(9,9,9), (9,9,9,9) **activation:** logistic, relu **solver:** lbfgs, adam	**hidden_layer_sizes:** (9,9,9) **activation:** logistic, relu **solver:** adam	0.923
2	**hidden_layer_sizes:** (3,3,3),(6,6,6),(9,9,9), (12,12,12) **activation:** logistic, relu **solver:** lbfgs, adam	**hidden_layer_sizes:** (9,9,9) **activation:** logistic, relu **solver:** adam	0.923

4. Model Performance

4.1. AUC-ROC & Gini

cross_val_score is used to get the key model performance matrices.

- AUROC (CV=5) is 0.923
- Gini (CV=5) is 0.847

4.2. Cut-off of 0.4

Based on the F1-Score the probability cut-off is decided

- Prob > 0.4 then 1
- Prob <= 0.4 then 0

5. Version

- **Python (Version: 3.7.6)**
- pandas (Version: 0.25.3)
- numpy (Version: 1.19.5)
- scikit-learn (Version: 0.22.1)
- matplotlib (Version: 3.1.3)
- seaborn (Version: 0.10.0)