Credit Card Default Analysis – Machine Learning Algorithms

Xinxie Wu, xinxiewu@gmai.com

Predicting

In banking, Credit Risk is a big issue; Banks use various techniques for **Credit Card Default Analysis**. This research applies **2-part methodology** into default analysis and prediction, based on the dataset of **Default of Credit Card Clients**.

In **baseline** approach, we train Logistic Regression, Naïve Bayes, Gaussian Discriminant Analysis, Decision Tree and SVM for both continuous and discrete features, with total accuracy of ~80% but imbalanced pos-neg gap, >45%.

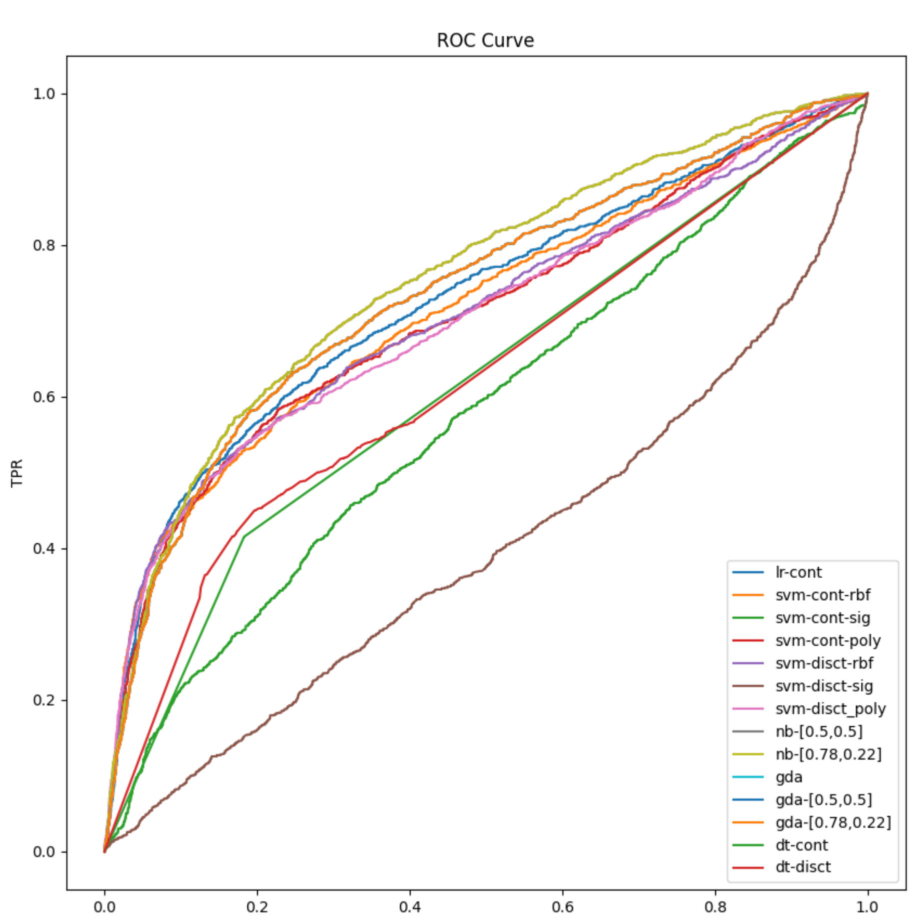
In **improvement** stage, SMOTE removes the imbalance gap, but brings overfitting; PCA/K-means refined the dataset and SVM's retraining results in 99% accuracy; Neural Network is trained and improves accuracy to 90.06%

Baseline Models

For baseline algorithms:

1. **Continuous**: SVM, with rbf kernel, achieves 81.97% accuracy; GDA reaches the highest AUC-ROC as 0.74, with F1 score 0.50.
2. **Discrete**: Naïve Bayes, with (0.78, 0.22) prior distribution, gains 80% accuracy; also, this NB reaches AUC-ROC as 0.76.
3. **Logistic Regression** shows the highest imbalance gap (73.5%)

Model	Prevalence	Total Acc	Neg Acc	Recall	Precision	F1-Score	AUC-ROC
NaiveBayes-discrete-[0.78, 0.22]	21.88%	80.22%	90.25%	44.40%	56.06%	49.55%	0.7546
LogisticReg-continuous	21.88%	80.97%	**97.06%**	**23.53%**	69.13%	35.11%	0.7269
SVM-continuous-rbf	21.88%	81.97%	95.63%	33.21%	68.02%	44.63%	0.715



Dataset & Features: Default of Credit Card Clients

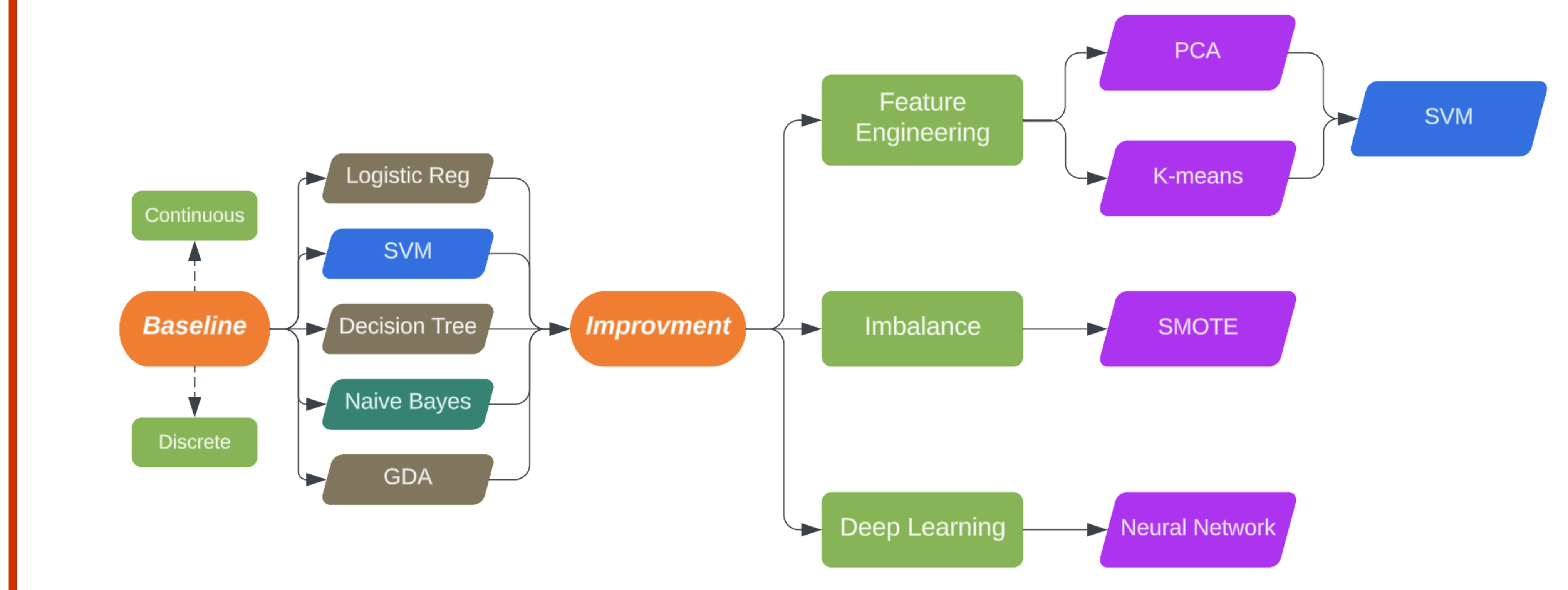
Attribute	Description	Type	Mean	Std	Min	25%	Median	75%	Max
Age	Age in years	Numeric	35	9	21	28	34	41	79
Limit_Bal	Given credits, family-based	Numeric	167,484	129,748	10,000	50,000	140,000	240,000	1,000,000
Bill_Amt_1	Bill statement of Sep, 2005	Numeric	51,223	73,636	(165,580)	3,559	22,382	67,091	964,511
Bill_Amt_2	Bill statement of Aug, 2005	Numeric	49,179	71,174	(69,777)	2,985	21,200	64,006	983,931
Bill_Amt_3	Bill statement of Jul, 2005	Numeric	47,013	69,349	(157,264)	2,666	20,089	60,165	1,664,089
Bill_Amt_4	Bill statement of Jun, 2005	Numeric	43,263	64,333	(170,000)	2,327	19,052	54,506	891,586
Bill_Amt_5	Bill statement of May, 2005	Numeric	40,311	60,797	(81,334)	1,763	18,105	50,191	927,171
Bill_Amt_6	Bill statement of Apr, 2005	Numeric	38,872	59,554	(339,603)	1,256	17,071	49,198	961,664
Pay_Amt_1	Previous payment of Sep, 2005	Numeric	5,664	16,563	-	1,000	2,100	5,006	873,552
Pay_Amt_2	Previous payment of Aug, 2005	Numeric	5,921	23,041	-	833	2,009	5,000	1,684,259
Pay_Amt_3	Previous payment of Jul, 2005	Numeric	5,226	17,607	-	390	1,800	4,505	896,040
Pay_Amt_4	Previous payment of Jun, 2005	Numeric	4,826	15,666	-	296	1,500	4,013	621,000
Pay_Amt_5	Previous payment of May, 2005	Numeric	4,799	15,278	-	253	1,500	4,032	426,529
Pay_Amt_6	Previous payment of Apr, 2005	Numeric	5,216	17,777	-	118	1,500	4,000	528,666

Dataset has 30k observations, 6,636 (22%) default; includes 23 attributes covering demographic and card historical information. All features are used and further analyzed by PCA & K-means.

EDA:

1. No Missing Value – Reasonable Values
2. Normalization & Discretization (9 Categories)
3. Correlation Matrix – Marriage & Age (0.41)
4. Training vs Testing: 80% / 20%

2-Part Methodology: Models & Workflow



Principal Component Analysis (23)

$$\underset{\emptyset_{11},\emptyset_{12},\ldots,\emptyset_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\emptyset_{j1}x_{ij}\right)^2$$

SVM re-training determines the optional # as 23

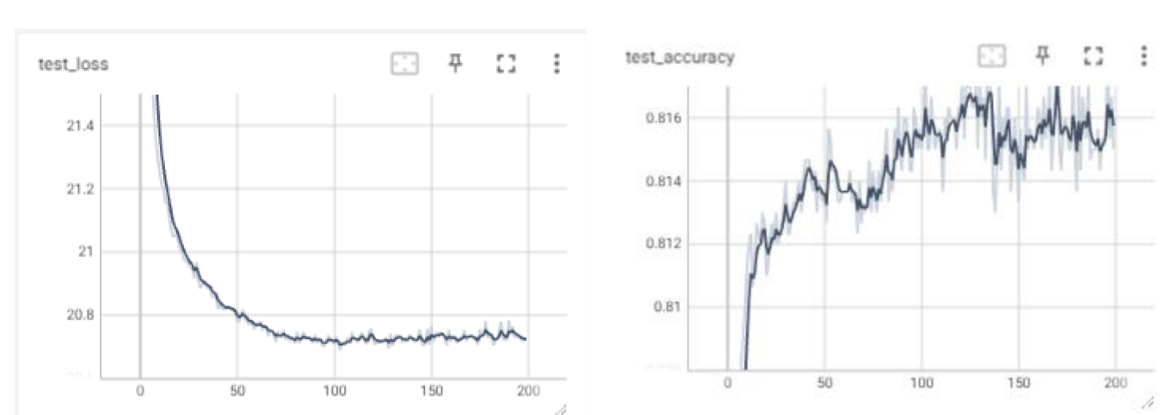
K-means (2)

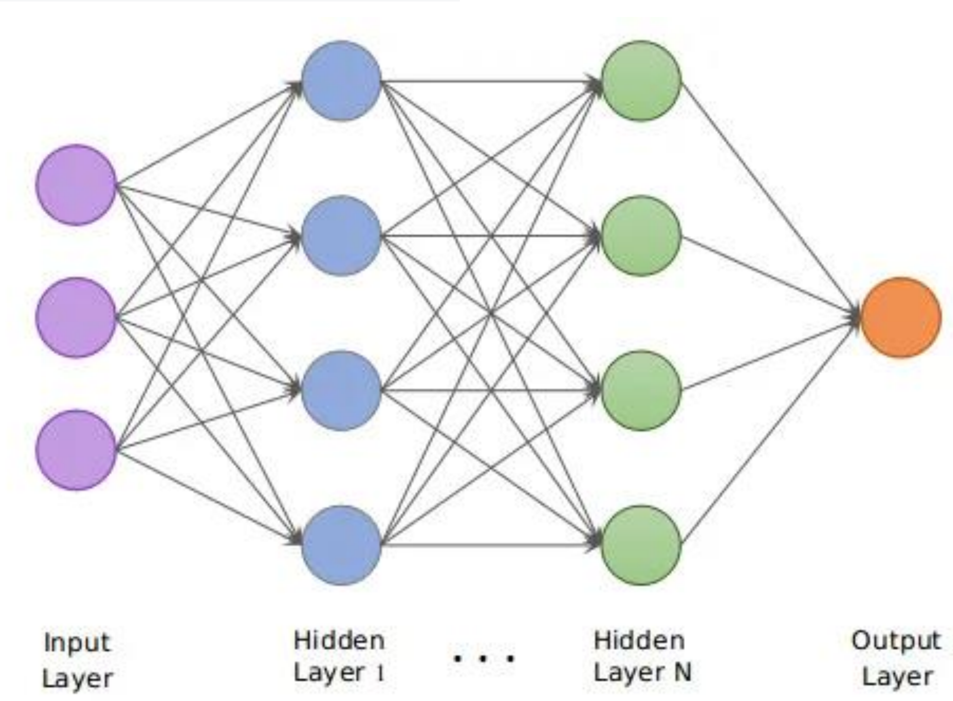
$$J=\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|x^{(i)}-\mu_j\right\|^2$$

Neural Network

Dataset is split into training (80%), validation (10%) and testing (10%).

Convolutional layer is NOT in this research.





Results & Discussion

Model	Total Acc	Nega Acc	Recall	Precision	Imbalance	F1-Score	AUC-ROC
Baseline Results							
SVM - Continuous - RBF Kernel	81.97%	95.63%	33.21%	68.02%	62.42%	44.63%	0.715
Naïve Bayes - Prior of [0.78, 0.22]	80.22%	90.25%	44.40%	56.06%	45.85%	49.55%	0.7546
Gaussian Discriminant Analysis	71.05%	72.75%	64.97%	40.05%	7.78%	49.55%	0.7371
Decision Tree - Discrete	75.62%	86.54%	36.63%	43.26%	49.91%	39.67%	0.6229
Logistic Reg. - Continuous	80.97%	97.06%	23.53%	69.13%	73.53%	35.11%	0.7269
Feature Engineering							
PCA(23) + KM(20729, 69.10%) = SVM	99.49%	99.90%	91.79%	97.94%	8.11%	94.76%	0.9998
Imbalance - SMOTE							
LogisticReg. - SMOTE	68.60%	69.40%	65.73%	37.57%	3.67%	47.81%	0.7306
Neural Network							
Neural Network - Linear, ReLU, Dropout	90.06%	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Results & Discussion:

1. Baseline models show total accuracy ~80%, but >45% negative-positive gap;
2. SMOTE removes the neg-pos gap, but brings overfitting (68.60%);
3. PCA returns an optimal number as 23, K-means removes 9,271 obs. Improved SVM accuracy 99.49%;
4. Neural network includes 0.8 dropout and gets 90.06%; no convolutional layer in this research.

Future Work

For the future work, **k-fold cross-validation** is under consideration since our research focused on 8/2 dataset split. Also, **neural networks** with more different number of layers/neurons need to be trained and compare the performance. Finally, SMOTE shows overfitting and so poor generalization ability; methods besides sampling, such as **kernel-based** and **cost-sensitive**, should be considered and tested.

References

[1] Liu, R.L. (2018) Machine Learning Approaches to Predict Default of Credit Card Clients. Modern Economy, 9, 1828-1838.

[2] I-Cheng Yeh, Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, Volume 36, Issue 2, Part 1, 2009, Pages 2473-2480, ISSN 0957-4174.

[3] Husejinovic, Admel and Kečo, Dino and Masetic, Zerina, Application of Machine Learning Algorithms in Credit Card Default Payment Prediction (October 1, 2018). A Husejinovic, D Keco, Z Masetic, International Journal of Scientific Research 7 (10), 425-426, 2018.