Naive Bayes vs Decision Trees: Application in Credit Score Ratings

Kiril Tsarvenkov

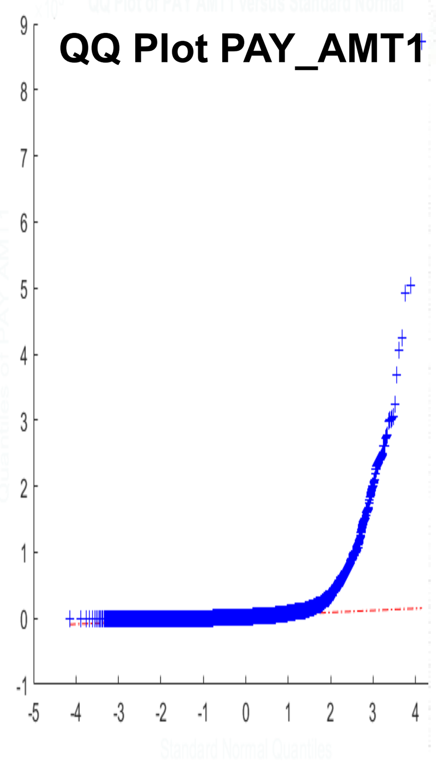
Description and Motivation of the Problem

- The aim of this poster is to compare Naive Bayes NB and Decision Tree DT algorithms for a binary classification task.
- The main task in to compare their ability to classify test/unseen data as either Default or Non Default (Payment) on client credit cards.
- The result of this study will be compared to the original results proposed by Yeh and Lien[1].

Initial Analysis of the Data

- Dataset: Default of Credit Card Clients from UCI
- The original dataset contains 23 attributes
  - 9 categorical and 14 numerical
- The response variable takes the value of 1 in case of a Default and 0 in case of a Non-Default (Payment)
- There are 30,000 observations without missing values.
- The numeric attributes BILLAMT_1 to 6 and PAY_AMT1 to 6 track the amount of the bill for a period of 6 months and the amount that was paid with respect to the loan.
- The distributions of the BILL variables are very similar, and the same applies to the PAY attributes
  - QQplots display typical BILL and PAY attributes which show serious deviations from Normality (red line)
- A formal test – Jarque-Bera JB – rejects the Normal/Gaussian assumption for all features

	MeanDef	StdDef	JBDef	MeanPav	StdPav	JBPav
AGE	35.73	9.69	1	35.42	9.08	1
LIMIT_BAL	130109.66	115378.54	1	178099.73	131628.36	1
BILL_AMT1	48509.16	73782.07	1	51994.23	73577.61	1
BILL_AMT2	47283.62	71651.03	1	49717.44	71029.95	1
BILL_AMT3	45181.60	68516.98	1	47533.37	69576.66	1
BILL_AMT4	42036.95	64351.08	1	43611.17	64324.80	1
BILL_AMT5	39540.19	61424.70	1	40530.45	60617.27	1
BILL_AMT6	38271.44	59579.67	1	39042.27	59547.02	1
PAY_AMT1	3397.04	9544.25	1	6307.34	18014.51	1
PAY_AMT2	3388.65	11737.99	1	6640.47	25302.26	1
PAY_AMT3	3367.35	12959.62	1	5753.50	18684.26	1
PAY_AMT4	3155.63	11191.97	1	5300.53	16689.78	1
PAY_AMT5	3219.14	11944.73	1	5248.22	16071.67	1
PAY_AMT6	3441.48	13464.01	1	5719.37	18792.95	1



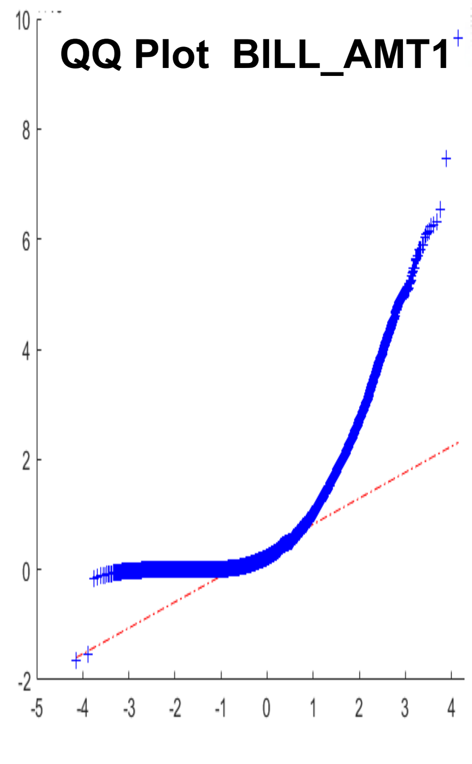


Figure 1. *Descriptive Statistics for the numerical attributes in case of Default and Non-Default (Payment). - Mean, Standard Deviation and Jarque Berra Test for Normal/Gaussian Distribution (at 5% confidence interval), displays 1 if the null hypothesis (normal distribution) is rejected and 0 if the null hypothesis is not reject*

Hypothesis Statement

- Yeh and Lien[1] report that Decision Trees outperform Naive Bayes in Credit Score rating and this study expects to achieve similar results
- Shorter Decision Trees (or less complex models) perform in a similar way compared to more complex models in terms of accuracy[2]
- Gaussian Naive Bayes should perform poorly on the test set due to the non-normal distribution of the features
- Both algorithms should achieve accuracy of more than 50% on the test set

Training and Evaluation Methodology

- The dataset is split into a training set of 20,000 observations and a test/unseen dataset of 10,000 observations.
- A fully grown DT is compared to shorter DT in order to control for the complexity of the model.
- Multinomial and Gaussian Naive Bayes models will be applied to the training set to evaluate which model performs most successfully. Binning will be applied to the numerical variables in order to convert them into categorical variables (referred to as multivariate multinomial).
- The algorithms will be compared in terms of classification accuracy and F-score
- The training set is left unbalanced in order to test the robustness of Decision Trees in unbalanced datasets as suggested by Mantovani, Rafael G., et al[3]

The Algorithms

Decision Trees DT

- Top down algorithm - places the attribute with the highest information gain at the top of the tree thus forming a root node[2]
- Successive nodes are created from the splits of the root node and from the attributes with the highest information gain until all attributes are classified *(Fig.2)*
- All training examples are used at each step
- Non-probabilistic classification algorithm based in entropy and information gain

Capabilities:

- ✓ The search hypothesis contains the target function
- ✓ Maintains a single hypothesis through the space of Decision Trees
- ✓ The search is robust to individual training errors and noisy data
- ✓ Computationally efficient and easy to implement

Limitations:

- ✗ Hill- climbing - the algorithm does not backtrack to evaluate earlier choices
- ✗ Does not find globally optimal solutions: converges only locally
- ✗ Does not determine how many alternative DTs are consistent with the data
- ✗ Preference/Search Bias for shorter trees and tends to overfit the data

Naive Bayes NB

- Probabilistic classification algorithm based on Bayes Theorem
- Assigns the most probable class using Maximum a Posterior rule given a set of attributes
- Assumes all attributes are independent of each other given the target values
- The posterior is calculated by the product of the prior and the class conditional probability for the individual attributes

Capabilities

- ✓ The classifier is optimal even when the independence condition does not hold[4]
- ✓ Optimal for learning conjuctions and disjunctions
- ✓ Computationally efficient and easy to implement

Limitations

- ✗ Overly simplistic (can be considered an advantage)
- ✗ Does not account for the interaction/dependencies between the features

Choice of Parameters and Experimental Results

Decision Trees DT - Parameters

- A basic fully grown DT is estimated in order to compare to other simpler models
- Shorter DT models are estimated with 4, 7, 14, 25, 50, 75 and 100 splits
- In selecting shorter trees,10 fold Cross Validation (CV) is implemented during training, in order to estimate a classification error across 10 folds for a given number of nodes, thus allowing for better estimation (does not involve selecting a model from the K-fold CV)

Results

1)The basic DT model produces a very large and complex tree with 3, 512 nodes
2)Shorter DTs with Cross Validation improve the estimates by approximately 8-9 %
3)The Cross Validated accuracy for DT models with 7 and 100 splits are very similar in magnitude, however, the former model is less complex compared to the latter *Fig.3*

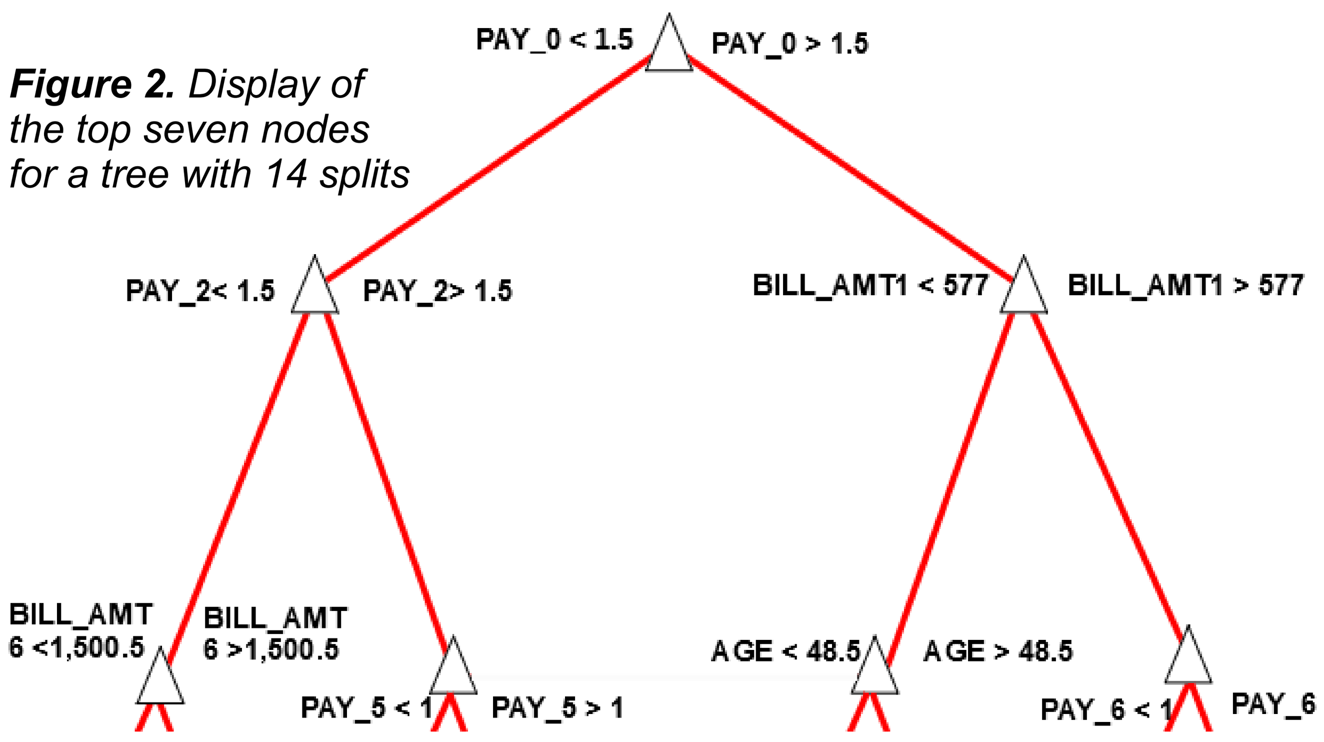
Naive Bayes NB - Parameters

- Prior is set according to the estimation results of the training set
- Given the non-normal characteristics of the attributes,multinomial estimations are carried out as alternatives to find the best model

Results

1) The Prior in the training set is estimated as: 77.21% Probability of Non-Default (Repayment) and 22.79% Probability of Default
2) Modeling the predictors as multinational increases the predictive accuracy of the algorithm compared to the Gaussian estimation
3) Gaussian estimation produces worthwhile results despite the non-normal characteristics of the attributes (*Table.1*)

Figure 2. Display of the top seven nodes for a tree with 14 splits



Decision Tree	Accuracy		Naive Bayes
Basic Model	74.38%	75.48%	Gaussian
Mdl 14	83.14%	79.50%	Multinomial
	F-score		
Basic Model	83.72%	83.70%	Gaussian
Mdl 14	90.11%	88.52%	Multinomial

Table.1 *Accuracy and F Measures for the Decision Tree – Basic model and Model 14 with 14 splits, and Naive Bayes – Gaussian and Multinomial models*

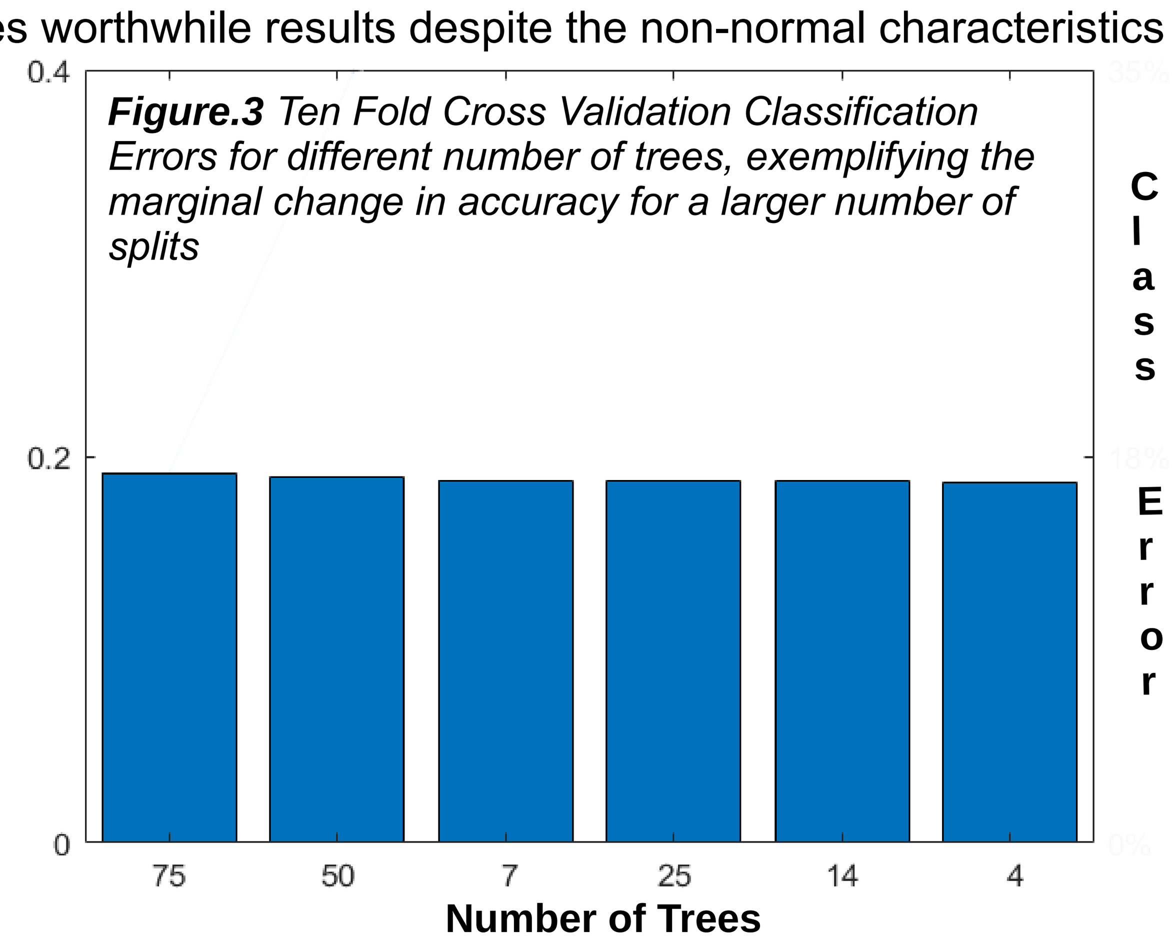


Figure.3 *Ten Fold Cross Validation Classification Errors for different number of trees, exemplifying the marginal change in accuracy for a larger number of splits*

Analysis and Critical Evaluation of the Results

- Decision Trees outperform Naive Bayes algorithm by a small margin thus the findings of this study are consistent with the result proposed by Yeh and Lien[1]
- Stopping the tree early and controlling for overfitting produces better estimates and a less complex model to analyze, which is consistent with the literature[2]
- Reducing the number of splits to 14 increases accuracy of DT by classifying more True Positives, however at the expense of more False Positives compared to the basic tree. Nonetheless, the proportional benefit is considerable in favor of the True Positives, which translates into approximately 6-7% improved accuracy and F scores *(Table 1)*.
- DT places PAY_0 (failure to meet a payment for a first time) attribute at the top of the tree *(Figure 2)* which means that this attribute contains the largest amount of information gain.
- Naive Bayes performs best for classification of Credit Scores under a Multinomial distribution assumption compared to Gaussian. The Multinomial assumption achieves this through an increase in the classification of Positive rates whether that be True positives or Negatives. This raises the suspicion that the algorithm is simply classifying the majority class to unseen observations which is a typical bias for an unbalanced dataset.
- DT with 14 splits outperforms Naive Bayes with a multinomial distribution assumption by a small margin. Nonetheless, the confusion matrices for the two models differ significantly. DT produce more Negative classification, whether that be True or False positives, whereas NB Multinomial is heavily skewed towards positive classification.
- The accuracy score for all models is lower compared to the F measure as the former exhibits bias in unbalanced datasets while the former is specifically designed for this apparatus

Lessons Learned and Future Work

- DT outperforms NB in classifying credit ratings. This is likely due to the unbalanced dataset that is being used. Future study can be conducted on balanced data.
- Assuming Gaussian distribution of non-Gaussian features in a NB model produces surprisingly accurate results.
- It would be interesting to experiment with approaches which address skewed distributions for Naive Bayes due to the heavy skew in the numerical variables in the dataset[5]
- Decision Trees and Naive Bayes are outperformed considerably by other algorithms[6], therefore future work will focus in Boosted Trees, Random Forests and SVMs.

References:

1. Yeh, I-Cheng, and 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, 2009, pp.73-80.
2. Mitchell, Thomas. Machine Learning. McGraw-Hill, London;New York;, 1997.
3. Mantovani, Rafael G., et al. Hyper-Parameter Tuning of a Decision Tree Induction Algorithm, IEEE, 2016, doi:10.1109/BRACIS.2016.018.
4. Domingos, P., & Pazzani, M. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. Proceedings of ICML '96
5. Kim, Sang-Bum, Hae-Chang Rim, and Heui-Seok Lim. A New Method of Parameter Estimation for Multinomial Naive Bayes Text Classifiers, ACM, 2002, doi:10.1145/564376.564459.
6. Caruana, Rich, and Alexandru Niculescu-Mizil. An Empirical Comparison of Supervised Learning Algorithms, ACM, 2006, doi:10.1145/1143844.1143865.