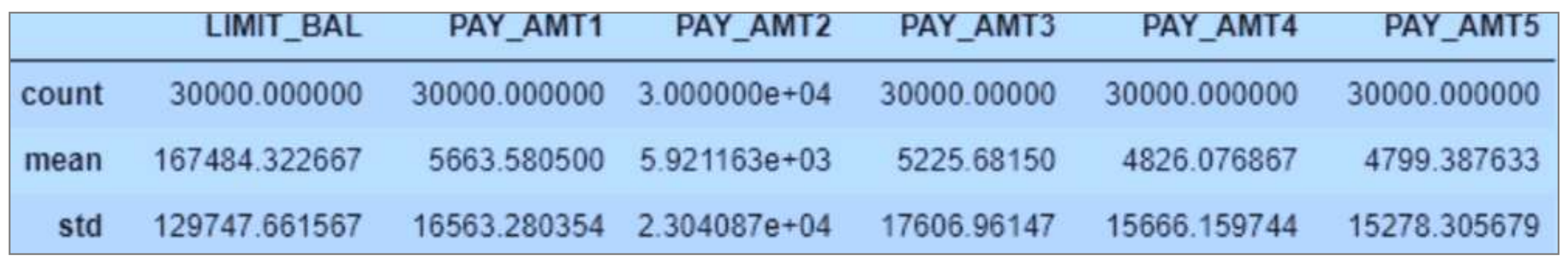
A Comparison of Naïve Bayes and Logistic Regression to Predict the Possibility of Clients Defaulting on Credit Card Payments

Divya Hebballi

Description and Motivation of the problem

- The aim is to compare the predictions of Naïve Bayes and Logistic Regression to find the possibility of customers defaulting on credit card payments in Taiwan.[1]
- In this binary classification problem, we will contrast our results to those obtained through previous research of I-Cheng Yeh and Che-hui Lien (2009).[1]

Initial analysis of dataset



	LIMIT_BAL	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5
count	30000.000000	30000.000000	3.000000e+04	30000.00000	30000.000000	30000.000000
mean	167484.322667	5663.580500	5.921163e+03	5225.68150	4826.076867	4799.387633
std	129747.661567	16563.280354	2.304087e+04	17606.96147	15666.159744	15278.305679

- Dataset: default of credit card clients Data Set from UCI.
- The original dataset has 24 features (10-categorical and 14 numeric).
- The features of the dataset were standardised using a Z-Score approach.
- We calculated the Variance Inflation Factor in order to deal with the multicollinearity in the dataset.
- We removed 9 features from the dataset which had a high Variance Inflation Factor namely marriage, age and Amounts of bill statements.

Two ML models with their pros and cons

Naïve Bayes

The technique calculates the probability of the classes given a set of feature values. The class is assigned based on calculation of the highest probability of a specific combination of features.

Pros:

- Prediction time is fast.
- Works well with high-dimensional dataset and limited training data.
- Easy to implement.

Cons:

- If the assumption of all features being independent does not hold true then the performance is weak.
- Cannot perform regression as it is a classification algorithm.
- Biased in nature because of independence assumption.

Logistic Regression

The approach uses a sigmoid function to model a binary dependent variable. The sigmoid function can take any real-valued number and map it into a value between 0 and 1 but never exactly at those limits.

Pros:

- Performs well on binary classification problems.
- Gives a measure of the relevance of the features.
- Easy to interpret, implement and very efficient to train.

Cons:

- Over fitting issues if the number of observations are lesser than the number of features.
- Predictions are discrete.

Hypothesis Statement

- I expect Logistic Regression to perform better. According to Baesens et al.[2] (2003) simple classifiers such as logistic regression perform very well for credit scoring.
- The previous research of I-Cheng Yeh and Che-hui Lien(2009)[1] however shows, that there are slight differences in performance between Naïve Bayes and Logistic Regression.

Description of choice of training and evaluation methodology

- Splitting the dataset into 70% train and 30% test dataset.
- Training the model on 21000 observations and testing on 9000 observations.
- Vary hyperparameters within each model to find optimum values.
- Evaluating using Accuracy, Precision, Recall, F1-score, ROC curves and execution time to identify the optimal model.

Choice of parameters and experimental Results

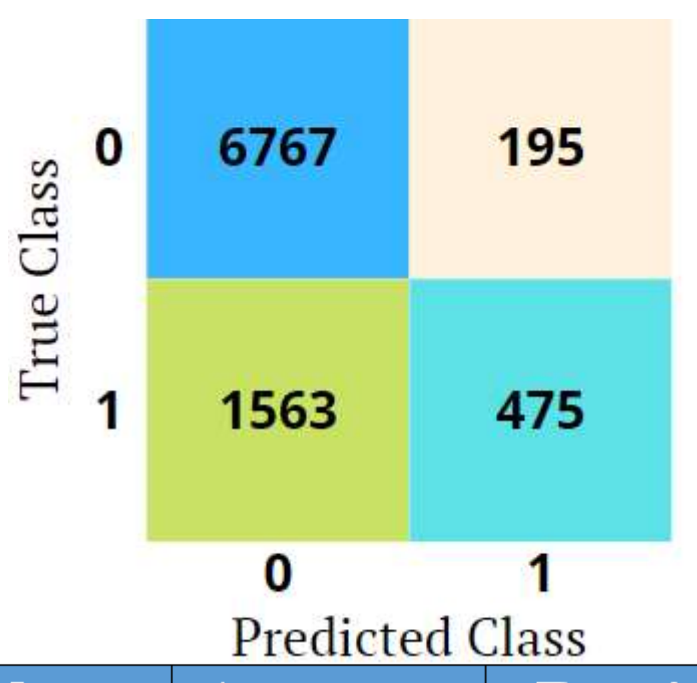
Logistic Regression

Parameters

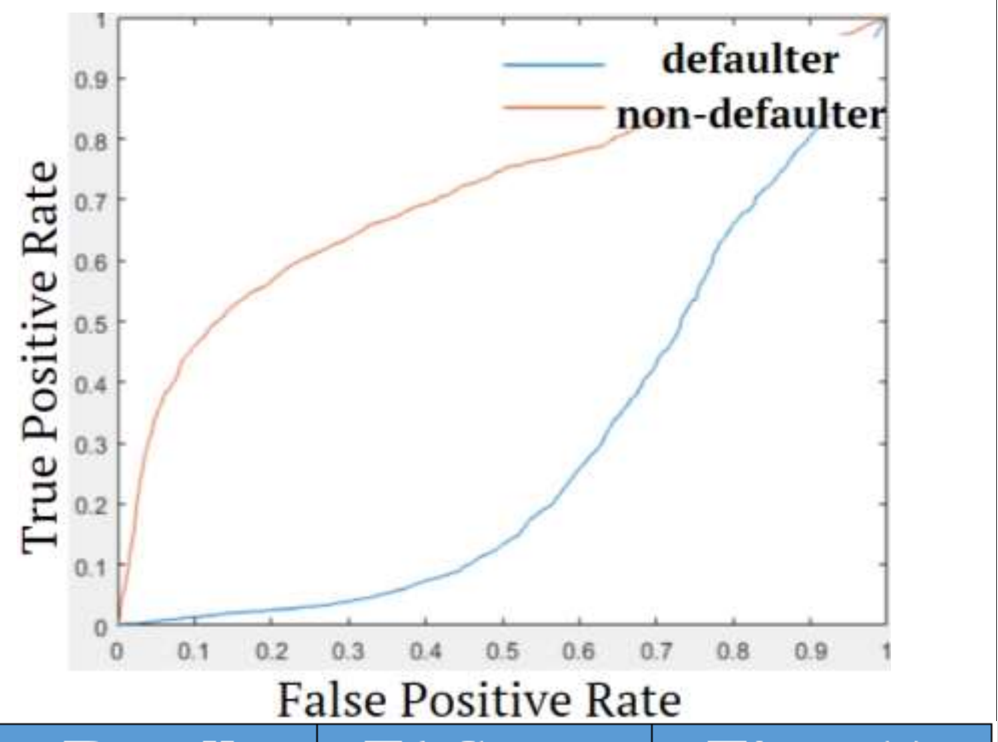
- Using Lasso Regularisation we try different values of lambda in order to penalise the features and optimise the model.
- Using lambda as the hyperparameter.
- Finding out the best value of lambda for finding the best features.

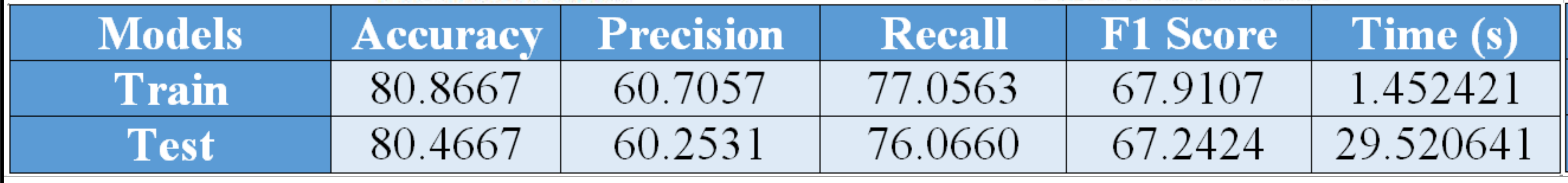
Main Experimental results

The best value of lambda resulted in a model with 5 features values as the input. The key elements of the 5 features were amount given credit and history of past monthly payment records.



	0	1
0	6767	195
1	1563	475





Models	Accuracy	Precision	Recall	F1 Score	Time (s)
Train	80.8667	60.7057	77.0563	67.9107	1.452421
Test	80.4667	60.2531	76.0660	67.2424	29.520641

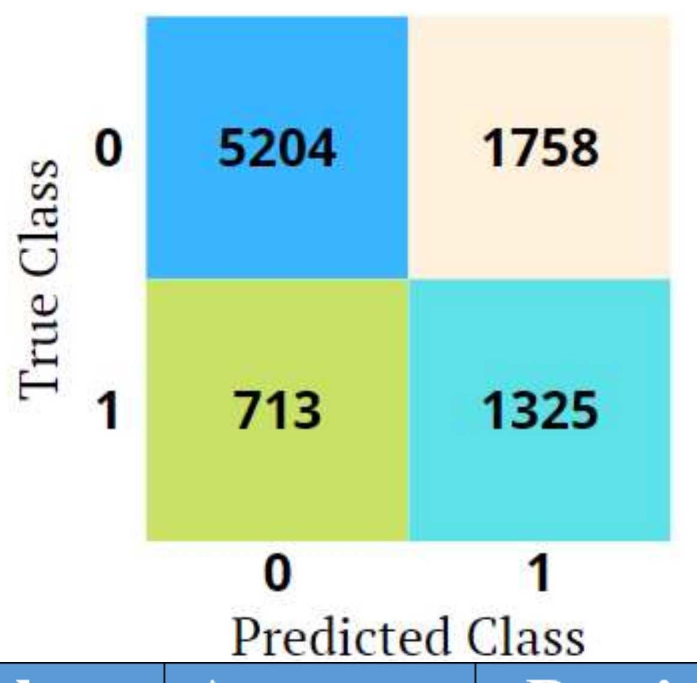
Naïve Bayes

Parameters

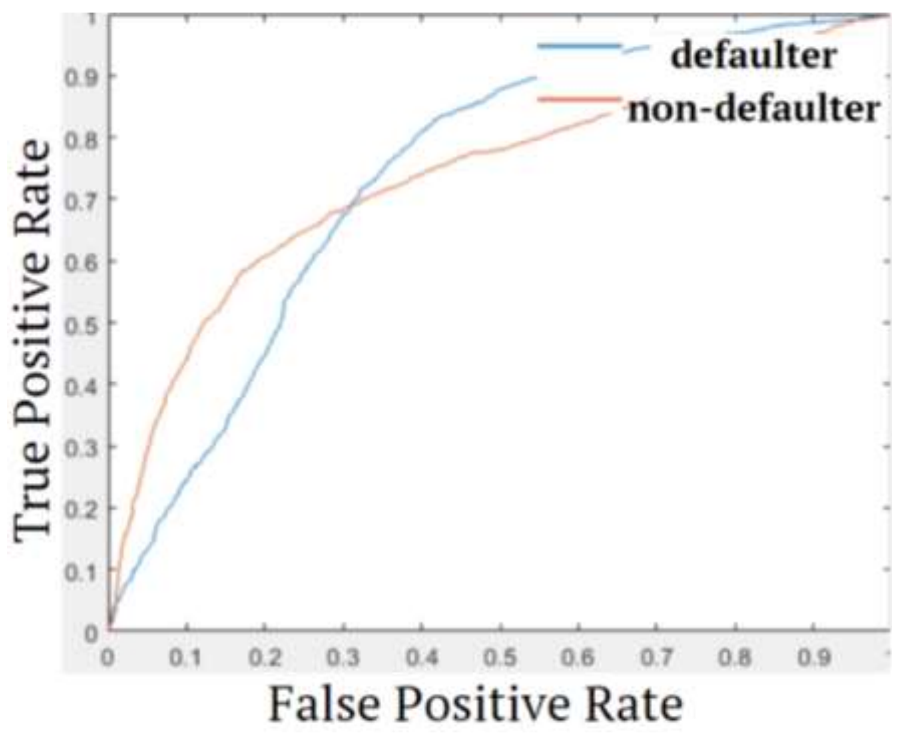
- Naïve Bayes can be optimised using the hyperparameters of distributions and priors.
- Manipulating priors, distributions to measure Recall, accuracy, precision and F1 score.

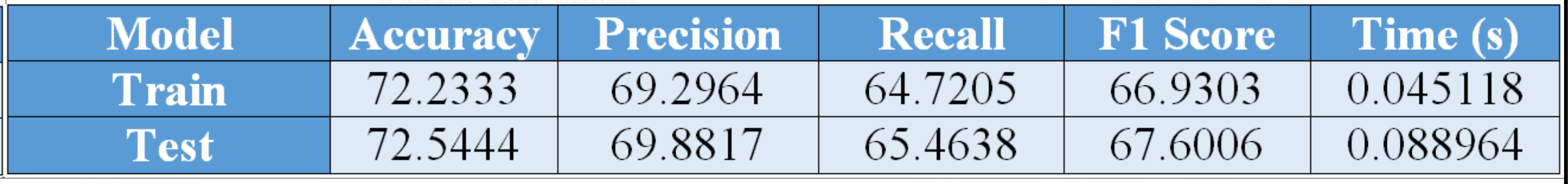
Main Experimental results

The normal distribution and empirical values of prior gave better accuracy and recall. Recall is a more powerful factor in understanding whether a customer will default.



	0	1
0	5204	1758
1	713	1325





Model	Accuracy	Precision	Recall	F1 Score	Time (s)
Train	72.2333	69.2964	64.7205	66.9303	0.045118
Test	72.5444	69.8817	65.4638	67.6006	0.088964

Analysis and critical evaluation

- After experimenting with undersampling of the data in order to remove the bias that would be introduced due to the imbalance in the classes of the dataset, the recall and accuracy dropped, so the untouched dataset seemed to be a better option in this case. However, the accuracy would not be the best way to measure the model performance owing to the class imbalance.
- A higher False Positive means that the model classifies more people as defaulters, and the goal of the business in our hypothetical case is to be more risk-adverse than precise, hence recall will hold more importance over precision and our aim is try to optimise the model to increase the recall.
- Logistic Regression can be optimised using feature engineering. We used the Lasso Regularisation method which encourages shrinking of coefficients to zero, and subsequently dropping those features. The experimental results however suggests that there wasn't any significant change in the performance when we incorporated all the features into the model. However, computationally the time taken for Lasso Regularisation was longer without achieving any significant improvement in results in our case.
- The Naïve Bayes classifier trains faster than Logistic Regression because it learns its parameters by explicitly calculating them, rather than learning its parameters by iteratively tweaking them to minimize a loss function. It introduces a bias in the process and our dataset doesn't seem to follow the bias well enough.
- The hyper parameter tuning in Naïve Bayes, which was done by trying different combinations of distributions and priors, did give us a better accuracy and recall. However, the experimental results suggests that in order to get better results with Naïve Bayes, additional data manipulation and feature selection helps to improve performance.

Future work

- The Dataset does have a class imbalance and after a failed attempt with under sampling it would be interesting to try other methods (SMOTE) to balance the dataset.
- More feature engineering and adding new features and given the current dataset could potentially be better predictors of credit card defaulters.
- Investigate an ensemble approach to have more improved performance.

1. Yeh, I. C., & Lien, C. H. (2009). 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients'. *Expert Systems with Applications*, 36(2), 2473-2480.
2. Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). 'Using neural network rule extraction and decision tables for credit-risk evaluation'. *Management Science*, 49(3), 312–329.·
3. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). 'Benchmarking state-of-the-art classification algorithms for credit scoring'. *Journal of the Operational Research Society*, 54(6), 627–635.
4. Xue, JH., Titterington, D.M. Comment on 'On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes'. *Neural Process Lett* 28, 169 .
5. Han, J., & Kamber, M. (2001). 'Data mining: Concepts and techniques'. San Fransisco: Morgan Kaufmann.