DSC3216/DBA3803 Predictive Analytics in Business

Final Project

Group 12

Members:

Name	Matriculation Number
Carissa Ying Geok Teng	A0205190R
Chanell Ng	A0203547J
Kuek Yan Ling	A0205292L

Table of Contents

1 Background

A home-seller wants to sell his house to one of the potential buyers who has made an offer. However, for the sale of his house to be successful, the buyer he has chosen should have the loan approved for the purchasal of the house. As the loan process is long and tedious, the home-seller would only know if the buyer's loan has been accepted or rejected some time after accepting the buyer's offer. Hence, the home-seller tries to make predictions on who will receive a successful loan application to decide who to sell to using analytics of past loan data.

A brief glance at the data can be seen below. The target is Loan_Status which indicates whether a loan has been approved (Y) or rejected (N).

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Applicant Income (thousands)	Coapplicant Income (thousands)
LP001003	Male	Yes	1	Graduate	No	4583	1508.0
LP001005	Male	Yes	0	Graduate	Yes	3000	0.0
LP002979	Male	Yes	3+	Graduate	No	4106	0.0

LoanAmount (thousands)	Loan_Amount_Term (months)	Credit_History	Property_Area	Loan_Status
128	360	1	Rural	N
66	360	1	Urban	Y
40	180	1	Rural	Y

Figure 1: Loan Data

The description of the data is as follows:

Variable and Definition	Loan_ID	Unique Loan ID of the loan applicant
	Gender	Gender of the loan applicant
	Married	Marriage status of the loan applicant
	Dependents	Number of dependents in the loan applicant's household (0,1,2,3+)
	Education	Whether the loan applicant is a University graduate
	Self_Employed	Whether the loan applicant is self-employed
	ApplicantIncome	Applicant income in thousands
	CoapplicantIncome	Co Applicant income in thousands (if there is a co-applicant)

	LoanAmount	Loan amount in thousands
	Loan_Amount_Term	Term of loan in months
	Credit_History	1: loan applicant's credit history meets set guidelines, 0: loan applicant's credit history does not meet set guidelines
	Property_Area	Residential area of the loan applicant (Urban / Semi Urban / Rural)
	Loan_Status	Y: Loan is approved, N: Loan is rejected

Figure 2: Variables and their respective definitions

2 Introduction

This report describes the use of various classification methods, in particular logistic regression, decision tree, random forest, gradient boosting and K nearest neighbours. Buyers are classified as having their loans approved or rejected. It is assumed that the home-seller wishes to target those who are able to have their loans approved in order to successfully sell his house, and to not target those who will have their loans rejected to reduce time wasted and the missed opportunities of other potential buyers who are able to obtain the necessary loan.

3 Dataset & Pre-processing

Since the dataset contains numerical, ordinal categorical, nominal categorical and binary categorical variables, each of these types of variables will be pre-processed in the following manner:

- Continuous variables: unchanged
- Ordinal categorical variables: one-hot encoding for logistic regression and k nearest neighbours, integer encoding otherwise
- Nominal categorical variables: one-hot encoding for all categories
- Binary categorical variables: binary dummy encoding (e.g. for Gender: "1" for male, "0" for female)

We have also identified the variable types of each feature:

- Continuous variables: LoanAmount, ApplicantIncome, CoapplicantIncome
- Ordinal categorical variables: Dependents, , Loan_Amount_Term
- Nominal categorical variables: Property_Area
- Binary categorical variables: Gender, Married, Education, Self_Employed, Credit_History

The Loan_ID is a unique identifier for each observation and is hence dropped during the preprocessing.
Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. Hence in our preprocessing steps for logistic regression and K-Nearest Neighbors, we have utilized the Pipeline library before calling GridSearchCV so that in every fold, standardisation will be called.

4 Scoring Metrics Used

To compare performance between different models, we cannot simply use misclassification rate as the dataset is imbalanced. With around 70% of training samples having their loans approved, there is a bias towards predicting that loans will be approved. Even with a higher accuracy, the model might perform poorly in other performance measures such as specificity. Hence, a better metric to use is area under the curve (AUC) for receiver operating characteristic (ROC) curve.

5 Models

The AUC Scores have been rounded off to 3 significant figures.

Model	**AUC Score (Training Set)**	**AUC Score (Validation Set)**
Logistic Regression	0.827	0.779
Simple Classification Tree	0.830	0.741
K Nearest Neighbours	0.786	0.734
Random Forest	0.882	0.800
Gradient Boosting	0.858	0.782

Figure 3: AUC Scores of Different Models

5.1 Comparison of Different Models

In general, all of the models have no severe concerns for overfitting as the AUC score of the training set is relatively similar to the AUC score of the validation set.

Comparing the AUC scores of the validation set of these 5 models, the models with the top 2 scores are random forest, followed by gradient boosting. However these two are considered more complex models. The other 3, Logistic Regression, Simple Classification Tree and K Nearest Neighbours are considered simpler models. Among the simpler models, Logistic Regression has the highest validation score. It is also worth noting the difference between these 2 AUC scores is very small so it is difficult to tell if this difference is due to luck or skill.

If it is due to luck, then there is not much to benefit from a complex model like random forest. Furthermore, since the difference in AUC scores is very small, a simpler model, like logistic regression, with greater interpretability is favoured more than a complex model like random forest.

Hence, the use of a simple or naive benchmark such as logistic regression, in this case, is justified as the gain from using a more complex model like random forest is not worth the additional complexity for a relatively small increase in AUC scores of the validation set (0.779 for Logistic Regression compared to 0.800 for Random Forest).

5.2 Logistic Regression

For logistic regression with regularization, GridSearchCV has been utilized to determine the optimal hyperparameters which are **{'model__C': 1.0, 'model__penalty': 'l1', 'model__solver': 'liblinear'}**. A relatively high AUC of **0.779** has been obtained on the validation set, indicating that the performance is good. **The validation set ROC curve and the covariate coefficients can be found below:**

Loan Amount	Applicant Income	Co Applicant Income	Dependents 0	Dependents 1	Dependents 2	Dependents 3+	Loan Amount Term 36
-0.0579	0.0000	-0.1586	-0.1311	0.0000	0.0000	0.0378	0.0000

Loan Amount Term 60	Loan Amount Term 84	Loan Amount Term 120	Loan Amount Term 180	Loan Amount Term 240	Loan Amount Term 300	Loan Amount Term 360	Loan Amount Term 480
0.0000	0.0000	0.0000	0.0011	-0.1811	-0.0528	0.0000	0.0000

Property Area Rural	Property Area Semiurban	Property Area Urban	Gender Male	Married Yes	Education Graduate	Self Employed Yes	Credit History 1.0
-0.2552	0.0692	0.0000	0.0000	0.1694	0.0000	0.0000	1.0110

Figure 4: Covariate Coefficients

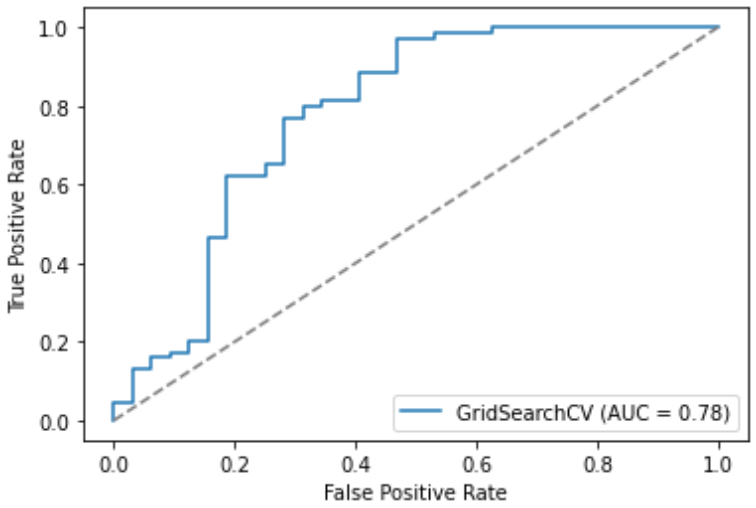

Figure 5: ROC-AUC on Validation Set

As L1 penalty is used, there is variable selection, resulting in a coefficient of 0 for many covariates such as Applicant Income. This helps to improve interpretability by identifying relevant features.

After choosing Logistic Regression as our preferred model, our team fit the test set to this model to obtain the ROC curve below. The AUC score of the Logistic Regression model for the test set is 0.7065, which is relatively high, demonstrating good test set performance.

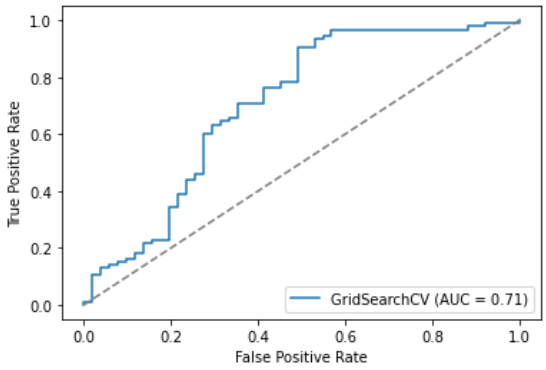

Figure 6: ROC-AUC on Test Set

6 Benefit Structures

The benefit structure allows for the determination of gain or loss for each predicted true negative (TN), false positive (FP), false negative (FN) and true positive (TP). The benefit score is determined by multiplying the confusion matrix with the corresponding weights in the benefit structure, such that the sum of products will return a scalar score, which can be normalised over the total number of samples.

We define TN, FP, FN and TP to be as follows:

- TN: Missing a buyer who ended up with a rejected loan status
- FP: Selling to a buyer who ended up with a rejected loan status
- FN: Missing a buyer who ended up with an approved loan status
- TP: Selling to buyer who ended up with an approved loan status

By adjusting the threshold, we can change how likely the model is to predict more positives or negatives. This will directly affect the confusion matrix, which will then affect the benefit score. For this analysis, we used the thresholds 0.1, 0.25, 0.5 and 0.75. Next, two benefit structures in two different scenarios are examined.

Scenario 1: When the home-seller is willing to wait and is not urgent to sell his house.
In this scenario, overestimation is better. Even if the chosen buyer to sell to might end up having a rejected loan status, the non-urgent home-seller can wait for more people to offer again. Hence, the cost of False Positives would be lower and the cost of False Negatives would be higher in this scenario. This is represented in the benefit structure below:

		Actual	
		Loan Approved	Loan Rejected
Predicted	Loan Approved	+100	-50
	Loan Rejected	-150	0

Cost of accepting a buyer: -10

Figure 7: Benefit Structure for Scenario 1

	Threshold	Benefit
0	0.10	11160
1	0.25	12660
2	0.50	12180
3	0.75	6540

Figure 8: Benefit for Different Probability Thresholds

With this benefit structure, the probability threshold with the highest benefit of 12660 is 25%. This is because the cost of FP is lower, while the cost of FN is higher, thus, the probability threshold is lower, resulting in more predictions for Loan Approved and fewer predictions for Loan Rejected.

Scenario 2: When the home-seller is urgent to sell his house.

In this scenario, underestimation is better. As the home-seller does not have the luxury of waiting, the chosen buyer to sell to has to have their loan approved. Hence, the cost of False Positives would be higher and the cost of False Negatives would be lower in this scenario. This is represented in the benefit structure below:

		Actual	
		Loan Approved	Loan Rejected
Predicted	Loan Approved	+100	-150
	Loan Rejected	-50	0

Cost of accepting a buyer = -10

Figure 9: Benefit Structure for Scenario 2

	Threshold	Benefit
0	0.10	4760
1	0.25	8760
2	0.50	8980
3	0.75	7440

Figure 10: Benefit for Different Probability Thresholds

With this benefit structure, the probability threshold with the highest benefit of 8980 is 50%. This is because the cost of FP is higher, while the cost of FN is lower, thus, the probability threshold is higher than that for Scenario 1, resulting in fewer predictions for Loan Approved and more predictions for Loan Rejected.