ID5059 Knowledge Discovery and Data Mining

Consumer Credit Risk Modelling with RBS

Team Cerro

20 April 2017

160025911, 160016606, 030007254, 160022313, 160025271, 160010979

Contents

1 Introduction

1.1 Executive summary

The purpose of this assignment was to develop a predictive credit risk model for Royal Bank of Scotland (RBS) based on a provided set of applicant data.

Initially, this report reviews a selection of literature discussing the most commonly used criteria and relevant issues relating to consumer credit risk modelling. Our approach to generating the most suitable model involved two iterative phases: firstly, data exploration and imputation, then predictive modelling and testing. We considered and implemented various ideas at both stages, including the use of various possible modelling techniques. The resulting models were compared based on their accuracy and area under the curve ("AUC") to find the one with the best predictive ability, whilst also considering other factors such as whether feedback on decisions can be obtained from the model. Section 4: Findings contains details and results for all considered models.

Our final model recommendation with an estimated predicted accuracy of 66.0% and AUC of 72.4% was a Penalized Regression Spline from data with MICE imputed missing values. Boosting and bagging also warrant a special mention as they too performed admirably. However, the transparency of the models became a deciding factor between the highest performing models.

1.2 Assumptions

While this practical was a useful exercise in solving a relatively complex decision-making problem for a real-world situation, one significant drawback of this approach was that we were unable to offset this real-world complexity with actual opportunities for communication with the Client. This lead to a number of ambiguities. As such, we had to make a number of key assumptions that would ordinarily be clarified by a Client. These assumptions are not necessarily limiting ones *per se*, but do impact the interpretability of the results in a real-world setting. The key assumptions adopted are summarised below:

Missing values in response variables

We assume that the PASS responses offer no possibility of knowledge discovery beyond using their covariates for imputation. It was initially ambiguous whether PASS should be interpreted as: (i) existing between GOOD and BAD on a scale of account performance (as implied during clarification lectures); (ii) worse than BAD (as implied in the assignment specification); (iii) not on a scale of account performance at all, as it is not strictly mutually exclusive with GOOD and BAD (which we have deduced to be the case).

In the training data, negative app_ids map 1:1 with a PASS in the response variable. Therefore we see no merit in attempting to predict it. Furthermore, in an email from RBS, we were told that "Application ID is not a relevant field for the students". It must follow that any variable that completely mirrors this

information must also be irrelevant. Whether an application is a "PASS" or not happens to be such a variable.

This insight from the Client is consistent with our interpretation. More specifically, a PASS implies a decision has already been made for this application based on some other criteria. Therefore is it not only impossible to use this application to predict probable performance of future applications (we do not know whether the application would have turned out to be GOOD or BAD), but it is also illogical attempting to simply recreate a decision-making process that exists elsewhere in the Bank at a previous stage of the application.

Thus, this assignment presents a binary classification problem. As such, it should be noted that it is not appropriate to test our model against rows containing missing data for the response variable. If any credit is to be given for predicting PASS responses, we opt to mark all test rows with negative app_ids as PASS, in line with the 100% correlation observed in the training data.

What constitutes a GOOD or BAD customer

The Bank's decision-making process for on-boarding new customers into overdraft products is interpreted to be a profit-maximising exercise within the constraints of certain risk parameters. However, simply ensuring the avoidance of loss is not a suitable strategy for profit-making. For an overdraft product, widespread under-usage of overdraft facilities essentially costs a lender money. For every credit line that a Bank offers a customer, it must withhold a certain amount of regulatory capital [15]. This is to give the Bank a "buffer" in case of some event that may cause widespread utilisation of credit lines or withdrawal of deposits. There is an opportunity cost attached to holding such regulatory capital aside for each product, as these funds could be offered to a more profitable debtor, invested in some part of the business, or given to shareholders through dividends or buybacks.

It is for this reason that customers who are in an extremely stable financial position and typically do not ever utilise their overdraft at any stage are not necessarily "good" for the Bank. However, at the other end of the spectrum is the lower credit-worthy individual who is at risk of being subject to bankruptcy proceedings where the Bank only receives partial repayment of the principal amount attached to their delinquent account. The problem underlying this exercise is in fact about striking a balance between these two aforementioned groups to find the happy-medium group who will utilise their overdrafts, pay interest, but also return the account to credit on a regular basis in line with their schedule of income. Therefore, in reality, optimal customer selection is a much more complicated problem than the one presented in our data.

We have made a simplifying assumption that this predictive model will be used solely for identifying whether an application is in danger of becoming delinquent or not, and that this is indicated by a non-zero and non-missing BAD response variable. Similarly, we assume that GOOD simply indicates an absence of BAD.

2 Research

2.1 Literature review

The following literature review addresses the most commonly discussed criteria to successfully build a predictive model in a consumer credit risk modelling context.

Establishing correlates / ratios

Quantitative data attributes can often be used to derive new features, such as ratios, in order to enhance the predictive ability of a model. Credit card balance-to-income ratio is a ratio being commonly used in the literature. Khandani et al. (2010) determined that customers with a high rate in credit card balance-to-income also had a significantly higher delinquency rate. Additionally, the authors calculated a ratio comparing current income with its historical levels. The resulting metric is an indicator for significant negative income shocks caused by unemployment. This is useful information, not only because it is directly linked to higher delinquency rates, but also because it acts a proxy for otherwise sensitive personal data which might fall under strict privacy protection laws. In the case of Khandani et al. (2010), readily available data on unemployment could not legally be used which made the ratio a valuable source of knowledge about the customers' employment statuses [3].

The use of customized ratios to improve predictive ability is something we tested on our models. This is described in more detail at the end of Section 3: Model Implementation. Three ratios were tested: credit card debt-to-income ratio, utilisation rate (measured by credit card balance over total credit card limits) and proportion of life employed (measured by number of years employed over age). In the end, we found that there was no change in predictive power provided by these ratios.

To impute or not to impute

Clean and valid data as the basis for developing a consumer credit model is of the utmost importance considering the financial impact an incorrect model can have for a large financial institution such as RBS. In general, a model to estimate the probability of default for a new applicant is based on internal data, which is often incomplete or lacking sufficient history. This can in turn lead to unreliable and incorrect scores [8]. In practice, when missing at random, the missing values have often simply been taken out (list-wise deletion) without trying to "use estimation algorithms to enter a value" [16], but this can lead to a significantly diminished data set as well as introduce bias for the outcome as pointed out by Florez-Lopez [8]. Furthermore, in the same empirical study run by Florez-Lopez (2010) using an Australian credit approval data set, the best

results in terms of unbiased and stable estimates and classification accuracy were achieved with maximum likelihood and multiple imputation techniques.

For the given RBS data set, we imputed missing values firstly using a basic mean/mode imputation and secondly with the MICE package, which utilises multiple imputation by chained equations. In line with the findings presented by Florez-Lopez (2010), the MICE imputation generally resulted in superior predictive models.

Feature selection / trimming

According to Thomas (2000), the presence of ten to twenty attributes makes up a robust scorecard. This means that characteristics that are i) only making a small impact on accepting or rejecting an applicant; ii) correlated/collinear; and iii) not stable over time; need to be removed. Ways of deducing which variables to eliminate include linear regression on subsets of characteristics and evaluating out variable importance using forward introduction or backward removal [16]. An example of such a variable in the RBS data set might be current account balance, which is constantly changing and therefore not stable. If a predictive model was overly reliant on such a variable, it might give a different result to the same applicant if they applied immediately before or after their salary was paid. This is not helpful in identifying the systematic component of the phenomenon under study. We carried out variable importance assessments using both PCA and random forests, which is explained in Section 3: Model Implementation.

Complementary data sources

Once credit bureaus came into existence, the information they held relating to the performance of a consumer with different lenders, electoral and legal records were utilised to supplement the internal data Banks were already using to generate a model for future applicants [16]. However, relying solely on this data could seriously impact the modelling outcome due to a potential lack of consistency in the information provided by the credit reporting agencies as well as a missing link to a customer's situation such as their employment or health status [5].

In any case, we were unable to adopt the approach of using complementary sources as the provided data was fictitious, and therefore did it contain any identifiable information from which to look up applicants. The possibility of using an automated program to repeatedly enter the data from each row into an "estimate my credit score"-style online service was considered, but believed to be both non-trivial to implement as well as not necessarily in the spirit of the assignment.

Regulatory environment

Machine learning techniques offer most organizations the opportunity to enhance their decision-making through data. Financial institutions hold a large

amount of rich data which can be used to model their customers' behaviour, from transactional information to descriptive attributes. However, despite the fact that these techniques offer opportunities for improvement across all areas of a bank such as RBS, there are regulatory headwinds that may prevent their widespread adoption. Regulations governing risk models vary on the types of decision they are aiding. Some regulations require absolute transparency and verification of models. An example of this is the modelling process used by Banks' corporate banking divisions to determine how much regulatory capital must be set aside for a particular risk exposure [2]. As a result, a black-box model would not be permitted to be used as a basis for this decision. Our modelling assignment is related to a purely commercial customer acceptance decision, and is not currently subject to such regulations as far as we are aware. However, it is possible that restrictions on how this process is carried out may exist in the future. This could be either a direct restriction, or an indirect one such as adding a requirement that the Bank give each customer specific feedback on why a particular decision was made. If there is no significant reduction in predictive accuracy, using white-box models would be better preparation for this eventuality [2].

Additionally, there are several legal restrictions on the very data that is permitted to be used to derive predictions for commercial decisions. This is due to both data security and privacy concerns. Namely, the Fair Credit Reporting Act led Khandani et al. (2010) to strip all health care, insurance, unemployment, government, treasury account and file age data from their feature vector [3]. We have assumed all the data provided for this assignment adheres to current regulatory rules and best practices.

Use of scorecards for black box models

Closely linked to regulatory restrictions on machine learning is the need to give customers feedback on the decisions which were made based on an automated model. This is relatively straight-forward with white-box models such as classification trees. Further insights can be provided for other types of white-box models by using scorecards on a characteristic-by-characteristic basis so customers can see which attributes were below par in relation to their application. Black-box models such as neural networks do not allow this assessment of why a certain decision was made [16]. As a result, our assessment criteria for model selection requires a significant premium in terms of accuracy in order to recommend a black-box model.

Popularity of each model type

The most common approach to default risk modelling is logistic regression, according to Thomas [16]. Due to the fact that logistic regression returns a probability for binary classifications, it is a good option for the RBS data set. It is also helpful for developing scorecards as the final credit score for the applicant can be broken down into scores for each predictor and hence provide individual

values for each attribute [12].

Another common predictive modelling strategy is to use classification trees to split the data repeatedly until it results in several groups of either good or bad customers. This method is attractive due to the ease with which it can be explained, as well as its compatibility with "rule-based systems" such as underwriting. Hand (2001), however, states that these models were not applied much in this context at the time of writing [11]. More recently, Khandani et al. (2010) argue that the use of generalized classification and regression tree models (CART) in the banking sector does gain a significant advantage from its interpretability while other black-box models are viewed with "suspicion and skepticism" [3, p. 20].

Galindo & Tamayo published another paper which supports the use of CART models in risk modelling. Besides the aforementioned advantages, they also saw a superior accuracy for CART compared to neural networks in a classification problem of mortgage loan data. The authors achieved an average error rate of 8.31% with CART models, whilst the best result of a neural net provided 11.00%. Both had been trained on a data set of 2000 records [10]. This is relatively small compared to the RBS data set.

Survival Analysis

Whilst the focus of credit risk modelling historically was to split customers into GOOD and BAD ones, survival analysis aims to predict when, not if, a customer might default. It is a commonly used technique in medical studies, biological research or social sciences for example [14]. Instead of excluding a customer who cannot be classified for some reason, the useful information of how long they have already been making repayments can still be included [1]. Additionally, survival analysis provides not only one but various probability of default estimates depending on different time frames for each applicant. Generally, information like marital status or home ownership are taken as fixed variables for credit risk modelling during an application, process, but these might change over time and significantly influence the financial situation of a customer. Survival analysis enables the inclusion of time-dependent variables as well as other economic factors e.g. interest rate [1].

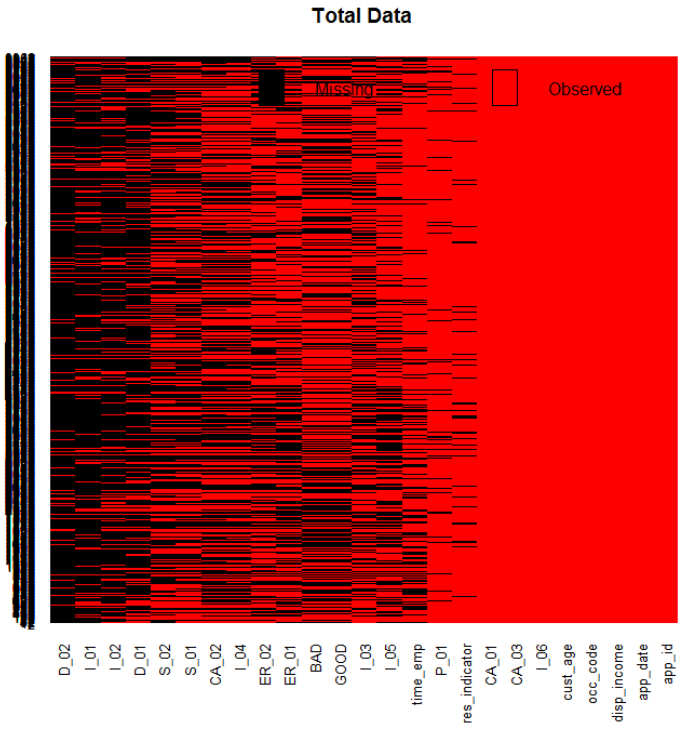
3 Implementation

3.1 Data exploration

The RBS data set used for modelling contains 9962 observations and 25 variables. The negative instances of variable S_02 (Applicant - Minimum number of months since credit search last 12 months) were assigned as NA together with

time_emp with values 99999. The number of remaining rows with complete observations in all 25 columns is 65. The "Amelia" package in R, graphically describes the "missingness" in the data below:

Figure 1: Missingness across explanatory variables



The variables above are ranked in terms of "missingness "with variable D_02 having 86.55% of the rows missing , closely followed by I_01 with 80.7% , while columns CA_01, CA_03, I_06, cust_age, occ_code, disp_income, app_date, app_id all have complete 9962 observations.

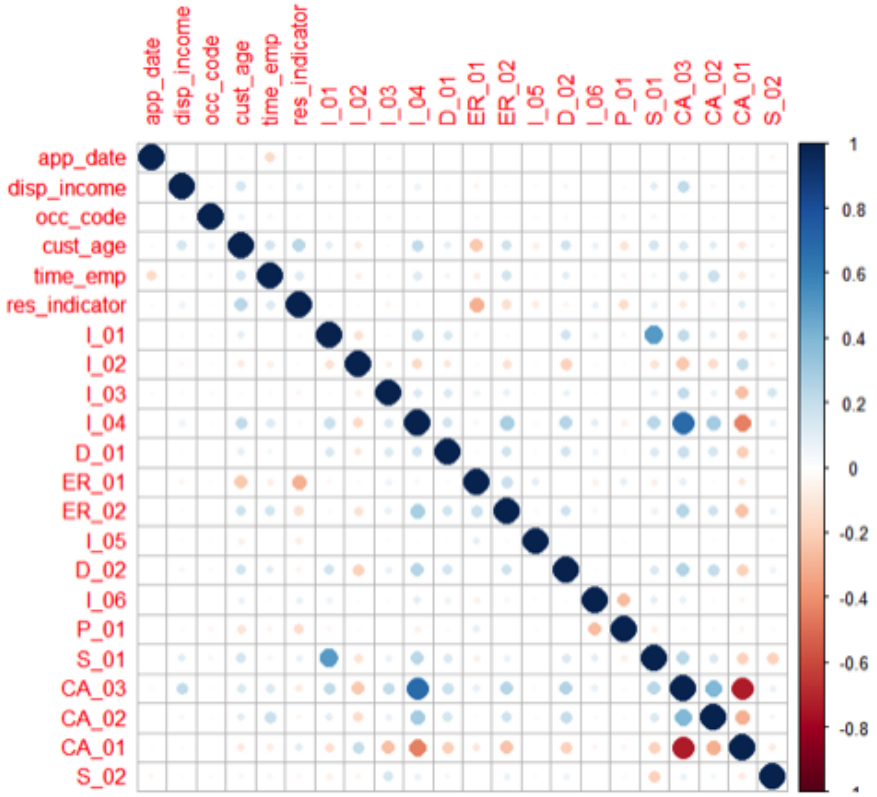
3.2 Treatment of missing values in covariates

Assessment of relative importance of covariates

By performing Principal Component Analysis ("PCA"), it is found that, based on the first and second principal component, the disposable income and P_01 "i.e. Average % of households with a live credit card" are the covariates which have the most impact on the outcome variable. The impact from disposable

income is positive, so it is less likely that a customer's account will turn "BAD" as disposable income increases. With P_01, on the other hand, it is more likely that customer's account will turn "BAD" as it increases.

Figure 2: Relationship between covariates



In this graph, each marker represents the relationship between two variables. The darker the shade of blue, the more positively correlated they are and the darker the shade of red, the stronger the negative correlation. The size of the circle additionally increases with a stronger correlation. This tells us that, for example, the covariates CA_03 (Number of live current accounts) and I_04 (Total number of accounts (including loans, cards, contracts, etc.)) are highly positively correlated. CA_03 and CA_01 (Worst status of all current accounts in the last 6 months), on the other hand, are negatively correlated. These insights can help with imputation and variable selection for the modelling process.

Imputation methods

We considered two imputation methods. The first involved using mean and mode imputation. This imputation was run across the whole data set, then split into two data sets: one where the response variable (GOOD/BAD) was complete (to then be used for model building); the other where it was incomplete. The second imputation approach involved using the R package MICE, which carries out multiple imputation. Different imputation methods can be specified for each data type, as well as the number of data sets to be imputed. The real challenge comes with deciding which predictors to specify for each variable. This information is passed to the MICE function as a parameter in the form of a predictor matrix. Initially, we tried our own customized approach, by trying to examine the correlation between the different variables. However, this was eventually abandoned in favour of using the inbuilt functionality of the MICE package, which can specify a predictor matrix itself. MICE suggests a predictor matrix by examining the correlation between variables, as well as other criteria.

MICE imputed five data sets, which is the default setting. We used the first of them for model training and testing. The different statistical methods used to impute each data type, which MICE requires to be predetermined for each covariate, were predictive mean matching for numeric variables and CART for categorical variables.

Specifying other methods of imputation, particularly for categorical variables, proved difficult. Often MICE would throw an error. One drawback of our imputation method was that we did not manage to create a strategy to investigate the quality of the imputation, but merely rely on the fact that most of the models saw a positive, albeit minor, effect in terms of accuracy.

3.3 Model implementation

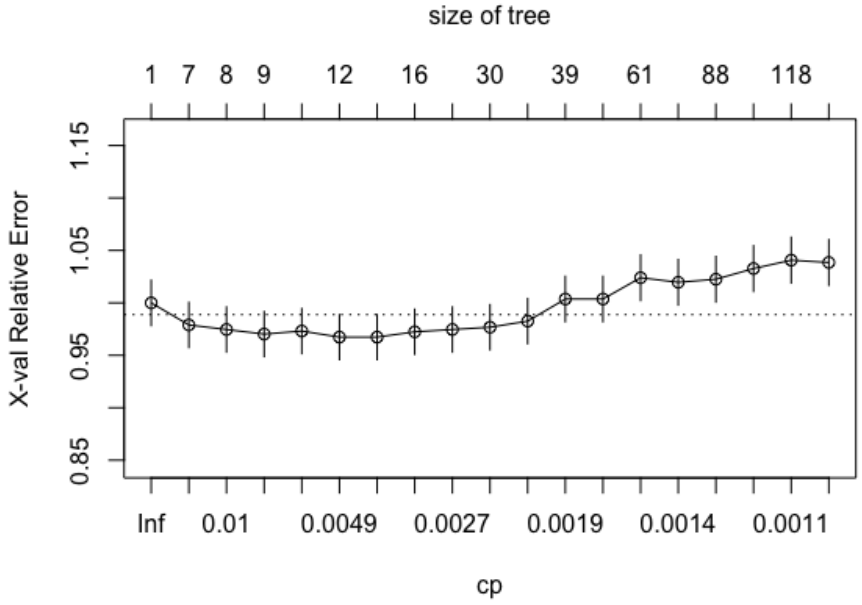
Trees

As discussed in the literature review [10], trees are regularly used in the financial sector as they suit the needs of the industry besides being accurate, parsimonious and feasible. Since predictive models are not run in isolation but as part of more complex models and bound to compliance guidelines in many cases they need to be transparent and interpretable [7]. This is beneficial not only to stakeholders but also to decision makers.

Our tree models were initially producing unrealistic outputs. This lead to some minor manual improvements in the data (such as assuming non-zero incomes below 10 as missing values) that led to a significant improvement in in the interpretability of the tree. Two methods were considered for constructing the trees: i) forward-selection; ii) backward-selection followed by pruning. We carried out both of these methods for the mean imputed data and the MICE imputed data. In both data sets, the forward-selection tree generated more parsimonious, and pragmatic results without compromising accuracy.

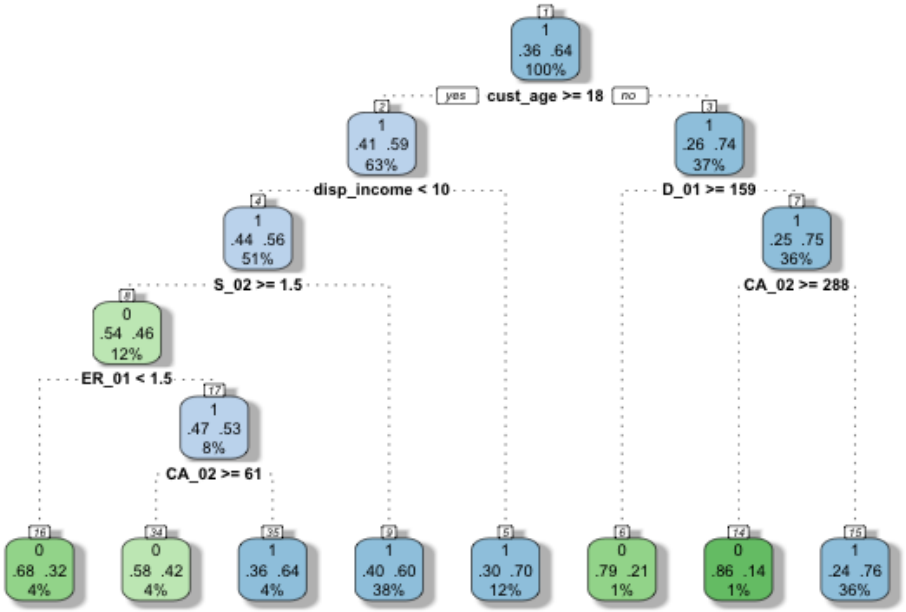
In the case of backward-selection, Figure 3 demonstrates an initial reduction in error from moving from a single node tree to a slightly larger, but still relatively parsimonious, tree of approximately seven or eight nodes. Beyond this there is little or no marginal gain, and eventually a marginal worsening.

Figure 3: cpplot - no significant reduction in error rate from 7 nodes onwards



The tree generated from the mean imputed data (Figure 4) was therefore chosen as our best tree model. It can be described as follows. The first node (and therefore, most important) variable is whether the applicant is over 18 or not. If they are not, there is a much larger probability that they will be classified as BAD (i.e. GOOD = 0). This depends on the results at the two following nodes on the right hand side of the tree: whether they have significant utility bill liabilities or a certain balance in their current accounts. Whilst the latter is a non-stable variable (i.e. it changes continuously), some applicants' balances will likely never go as low as this threshold, and this node helps identify that group regardless of their monthly cash flow situation. If, on the other hand, the applicant is over 18, their disposable income is assessed next. If this is over 10 (which we interpret as a quasi-check for 'non-zero' or not), they are classified as GOOD. This suggests a relatively un-strict criteria for product approval in the training data — specifically that the applicant be over 18 and have some positive amount of disposable income. If disposable income is below 10, then

Figure 4: Tree from mean imputed data



it is still possible to be classified as GOOD, if certain further conditions are fulfilled. Those conditions are: i) the applicant has had a credit search in the last 1.5 months (which is a peculiar result); OR ii) the customer has been at their current residence for a certain period (which could be interpreted as a proxy for whether they are a renter as opposed to the more stable situation of being a homeowner or living with parents), and they have a current account balance over a certain amount.

Boosting and Bagging

The *adabag* package was applied to implement boosting and bagging models for this assignment. Both of these model types were validated with 10-fold cross validation. For bagging, 50 trees were used in the ensemble. For boosting, 10 trees were used in the ensemble, all of which were stumps.

The strong performance of boosting and bagging overall is not surprising. As described in the literature by Khandani et al. (2010), boosting is commonly used to predict credit-default data. Boosting applies weights to scarcer observations more heavily than common ones [9]. Due to the typically highly skewed distribution of good and bad realizations, training a boosting model with unequal weights can lead to increased predictive power compared to standard CART [3].

Random Forest

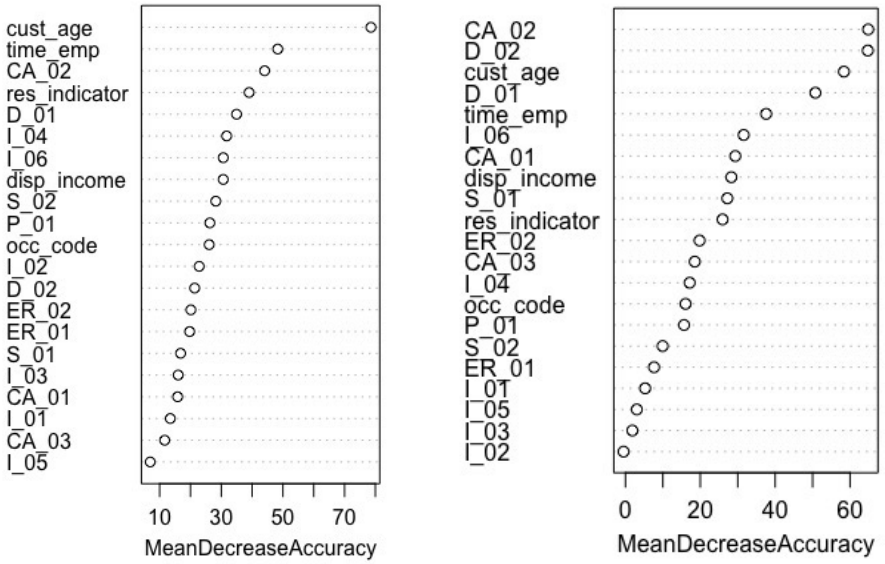
The R package RandomForest has been used as a method of prediction. As a random forest model cannot handle missing values, data needs to be imputed before it can be grown. Both imputation methods have been applied. Before feature selection the MICE-imputed data set returned a out-of-bag ("OOB") estimate of error rate of 30.21%. The mean-imputed data set resulted in a 2.69% higher error rate. As laid out by Breiman (1996), this value is sufficient as a measure of accuracy and no cross validation is required [6, pp. 123-140].

In an attempt to optimize the result, feature selection was also performed. The "Boruta" package was used as a non-traditional approach to feature selection. These algorithms aim for a minimal optimal subset of features by recursively removing features in each iteration which did not perform well in the process. This decision can be based on Mean Decrease in Accuracy ("MDA") or Mean Decrease in Impurity ("MDI") which are provided by the RandomForest package. Since the OOB estimate of error rate did not significantly decrease by this method the Boruta package has been applied.

In contrast to other selection algorithms, Boruta identifies relevant attributes which allow an understanding of the mechanism behind it instead of keeping them completely black-boxed. The algorithm uses a wrapper approach built around a random forest to compare the relevance of the real features to that of the random probes [4].

Unfortunately, applying this advanced feature selection method did not produce a better result than the traditional method. Boruta returned l_03 (Age in month of most recently opened account) and I_05 household (total number of accounts) as un-important, but these had already been identified as having a low mean decrease accuracy from regular iterative feature selection.

Figure 5: Feature Importance



Mean imputation MICE imputation

Naïve Bayes

A Naive Bayes Classifier ("NBC") model was run using all the variables in the data set. Initially, it was using the unimputed data set. One advantage of the NBC is its ability to handle missing values. This makes it an appropriate model to attempt for the RBS data set, given the level of "missingness" previously identified. It is highly likely that the key assumption of independence is not satisfied, although NBC has been noted to perform well even with this assumption violated [17]. The NBC model was run for both types of imputation considered (mean/median and MICE package).

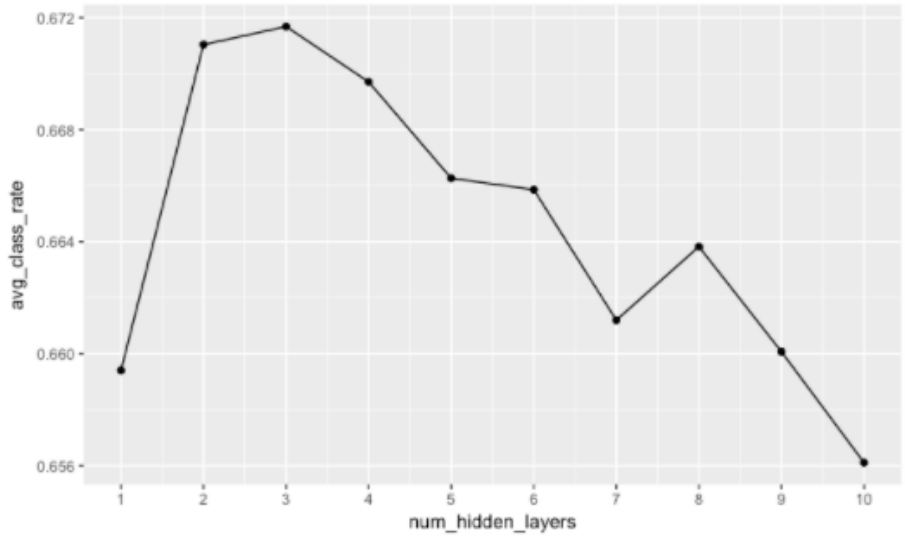
Neural Network

A neural network model was created and refined using the nnet package. During initial attempts, the networks would not converge when containing more than one hidden layer. The reason for this turned out to be significant, not only for the modelling of neural networks, but all models. When the data was initially loaded into R, RStudio mistakenly typed some Integer columns into Factor columns made up of many levels. Not only did this not allow the neural network to converge, but it also negated the predictive power of those covariates for all models (particularly trees - which could only split on a binary yes or no for each factor level, rather than incorporate inequality booleans at nodes concerning these "should-be" integer co-variates). The fix for this was to manually force

R to interpret each level as a numeric for the nine incorrectly typed covariates. Another alteration, required only for neural networks, was to normalize the data to between 0 and 1. This was achieved through a simple manipulation of subtracting the minimum value and dividing by the range. It was also necessary to remove the correctly identified factor variables (res_indicator and occ_code) for the package to run beyond the most simple level of complexity.

The remaining steps to complete the neural network were boilerplate - partitioning the data and selecting a level of complexity (number of hidden layers). No additional configuration to the default neural network were implemented, as the majority of this project was carried out before ID5059 lectures covered these possibilities. The results of the neural network models indicated a benefit from increasing complexity beyond one hidden layer to two, but beyond that increasing complexity reduced the accuracy of the model. This is likely due to the fact that the network was required to connect a non-trivial number of covariates.

Figure 6: Accuracy of Neural Network vs number of hidden layers



Logistic Regression

Logistic Regression models allow the probability of a customer turning BAD to change non-linearly with the covariate values. The probabilities are naturally bounded by zero and one and return predictions inside this range, which can therefore also be interpreted as a probability of positive classification themselves. This is beneficial to a financial institution as it allows for easy adjustment of

models in light of a change in risk appetite (e.g. a desire to on-board a higher number of new customers at the cost of slightly increased risk). Our logistic models are fitted using the glm function and by specifying various link functions (logit, clogclog). The predictive power of the model is assessed using a confusion matrix as the input data is binary.

Penalized Regression Splines

Under Generalized Additive Models ("GAM"), penalized regression splines can be used to model individual covariates. Splines are desirable to absorb and reflect sophisticated wiggly covariate behavior. Using penalized splines allows a reduction in number of decisions required for modelling. It is most commonly used among different types of splines. It uses 10 interior, equally-spaced knots for each covariate by default. Using the 2D penalized regression splines will allow the 2D behavior to be captured. The model used is considering interaction between customer age and disposable income with relation to residential indicator. Among many trials of different covariates, this proved to improve predictability. Other covariates are included when it is sensible and proved to increase model predictability.

3.4 Custom metrics

In accordance with information discovered during the literature review, we created a number of metrics to see if this increased the predictive ability of our models. Khandani et al. (2010) report that a debt-to-income ratio "greatly enhances the predictive power" of their model [3]. In line with this approach, we created the following metrics:

Table 1: List of custom metrics

Metric	Abbv.	Formula
Debt-to-income	DTI	D_02 / disp_income
Utilisation rate	UR	D_02 / I_02
Proportion of life employed	PLE	time_emp / age

These ratios were tested in both the Penalized Regression Spline ("PRS") model (which turns out to be our most accurate model so far in terms of AUC), as well as a random forest, which is commonly used in measuring importance of variables. In the case of PRS, none of these ratios affected the predictive ability of the model. The same was true in the Random Forest, although PLE and UR did rank quite highly in an analysis of Mean Decrease in Accuracy ("MDA"). However, the two constituent variables for each ratio also rank highly. Therefore,

the ratios are interpreted as adding no additional predictive power. Given the scale of missing values, it was also deemed preferable simply to include both constituent variables outside the ratio in case one of them is missing.

4 Findings

4.1 Classification accuracy results

Table 2: Classification accuracy results

Model	Opacity	Accuracy %		AUC %	
		Mean imp.	MICE imp.	Mean imp.	MICE imp.
CART	White-box	66.0	67.4	54.3	54.2
Random Forest	Black-box	66.7	66.6	56.8	56.6
Logistic Reg.	White-box	61.1	63.7	65.3	69.5
Naïve Bayes	Black-box	60.0	60.0	66.0	66.0
Boosting	Black-box	66.6	69.6	68.2	**71.8**
Bagging	Black-box	66.0	67.0	72.2	**73.0**
Neural Network	Black-box	65.9	65.5	50.7	53.0
Pen. Reg. Spline	White-box	66.5	66.0	66.5	**72.4**

Based on the above results, both in terms of accuracy and opacity, the Penalized Regression Spline ("PRS") is a suitable model to use for this decision-making problem. This was the model submitted at the "halfway" stage for accuracy feedback. In addition to its strong performance in terms of AUC, it is also a white-box model. This is the only distinguishing factor between our PRS model and boosting and bagging models, which came a close joint second. We see transparency as an important factor given the domain of this data, and this is also consistent with the literature we found for credit-default type models [2] [3] Whilst both AUC and accuracy are shown as measures of predictive ability, we attach more weight to AUC, given that the data is slightly biased by the large number of GOOD outcomes relative to BAD (3773 vs. 2040).

The feedback we received was as follows:

N = 4,981
You predicted 2,739 as good

Actual goods = 1,917
True good predictions = 1,827

We interpret this as an 80% accuracy rate based on the assumption that all PASS rows were predicted correctly if there were any in the test data (due to the fact that they are identifiable as having a negative app_id, and that no row with a negative app_id exists with a classification of GOOD or BAD):

Table 3: 3x3 Confusion Matrix

		Act.			
		GOOD	BAD	PASS	TOTAL
Pred.	GOOD	1,827	912	0	2,739
	BAD	90	2,152 - *np*	0	2,242 - *np*
	PASS	0	0	*np*	*np*
	TOTAL	1,917	3,064 - *np*	*np*	4,981

Where np = number of PASSES in test data

We calculate our accuracy rate as:

(1,827 + 2,152 - np + np) / 4,981
= (1,827 + 2,152) / 4,981
= 79.9%

The same result is calculated in the 2x2 confusion matrix (i.e. no PASS rows in test data):

Table 4: 2x2 Confusion Matrix

		Act.		
		GOOD	BAD	TOTAL
Pred.	GOOD	1,827	912	2,739
	BAD	90	2,152	2,242
	TOTAL	1,917	3,064	4,981

Where number of PASSES in test data = 0

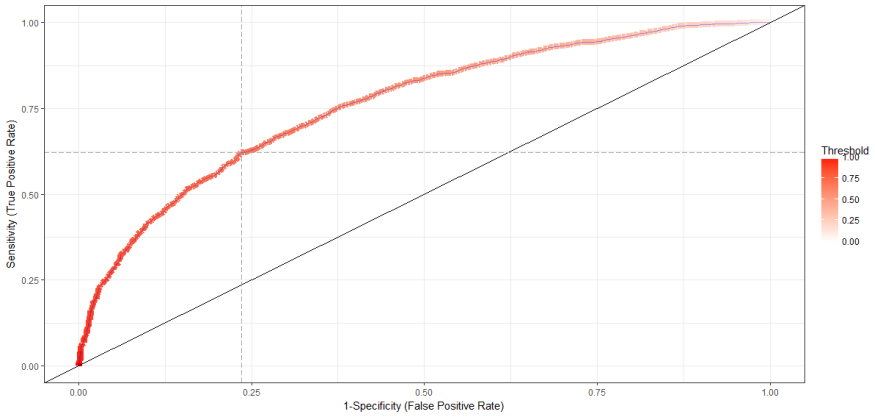
(1,827 + 2,152) / 4,981
= 79.9%

The only case where this accuracy rate is unknown, is where there are some PASS rows in the test data set with non-negative app_id. It was clarified before submission that there were no GOODs or BADs with negative app_ids in the test data set, therefore our accuracy of 80% is upheld at the halfway stage. This is interpreted as a good result, and therefore we opted not to make significant changes to our model as a result of this feedback.

4.2 Model uncertainty

As with any predictive modelling exercise, there are bound to be mis-classifications, particularly for cases whose predicted probability of positive classification are near the classification threshold. Our final penalized regression model classifies results greater or equal than 0.688 as GOOD = 1, and below as GOOD = 0. This threshold was chosen on the basis of it being the optimal point on the trade-off between specificity and sensitivity. The optimal point was assumed to be the point nearest the upper left corner, in the absence of any additional information from RBS regarding any preference for capturing as many true positives as possible vs. minimizing false positives.

Figure 7: Penalized Regression ROC curve



Having a hard threshold leads to very similar border line cases being classed differently even though they are very similar. We considered setting a softer threshold around 0.688, within which the model identifies the outcome as ambiguous. Whether this is useful or not depends on the real-world situation in which this will be used. We have assumed that this model acts as a "decision-aiding" model for a member of in-branch bank staff, and therefore can give its

output as a supportive opinion rather than a definitive decision. Maintaining the use of a hard decision-making threshold for similar classifications was consistent with literature we found such as Lessman (2015 [13]. In addition, the domain of this prediction is not of the nature of being life-critical or anything of the sort, so there is no explicit intolerance for classifying borderline cases without any additional flagging.

5 Conclusion

5.1 Summary and recommendations

In conclusion, we have implemented, validated and selected models from a menu of possibilities, and select a Penalized Regression Model as our best predictor of GOOD/BAD classification for the RBS data. From internal validation we estimated our predictive accuracy as 72.4%. Boosting and bagging models also performed very well, but these were not favoured due to the fact that it is more difficult to extract meaningful feedback concerning the reason behind each classification. This is perceived to be an important factor for RBS, whether enforced on them by regulators or otherwise in terms of their main priority: "to serve customers well" [15]. The strong performance of these models is consistent with the literature we found regarding predictive models in a similar domain.

The dataset contains a large number of missing values. In order to maximise our predictive accuracy we imputed the missing values, first using a mean/mode imputation, and secondly using the MICE package. We found the latter to have a better impact on our predictions.

This assignment hinged on several key assumptions which were not particularly clear, due to the lack of communication possible with the client. Whilst this is understandable, it may impact the interpretability of our results. As such, our recommendations should be considered alongside the assumptions set out in Section 1. Most importantly, due to the way we have treated instances of PASS in the outcome variable, it is not appropriate to test our final model against rows containing missing data for the response variable. If any credit is to be given for predicting PASS responses, we opt to mark all test rows with negative app_ids as PASS, in line with the 100% correlation observed in the training data.

Word count: 5,822

References

[1] Dirick et al. *Using Survival Analysis to Model Time to Default*. Jan. 2016. URL: http://www.dataminingapps.com/2016/01/using-survival-analysis-to-model-time-to-default/.

[2] Härle et al. "The future of bank risk management". In: *McKinsey Working Papers on Risk* (2015).

[3] Khandani et al. "Consumer credit-risk models via machine-learning algorithms". In: *Journal of Banking and Finance* 34.11 (2010).

[4] Kursa et al. "Feature selection with the Boruta package". In: *Journal of Statistical Learning* 36.11 (2010).

[5] Robert B Avery, Paul S Calem, and Glenn B Canner. "Consumer credit scoring: do situational circumstances matter?" In: *Journal of Banking & Finance* 28.4 (2004), pp. 835–856.

[6] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996).

[7] J Elder and Daryl Pregibon. "A statistical perspective on KDD". In: *Advances in knowledge discovery and data mining* (1996), pp. 83–116.

[8] R Florez-Lopez. "Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data". In: *Journal of the Operational Research Society* 61.3 (2010), pp. 486–501.

[9] Yoav Freund, Robert E Schapire, et al. "Experiments with a new boosting algorithm". In: *icml*. Vol. 96. 1996, pp. 148–156.

[10] Jorge Galindo and Pablo Tamayo. "Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications". In: *Computational Economics* 15.1 (2000), pp. 107–143.

[11] David J Hand. "Modelling consumer credit risk." In: *IMA Journal of Management mathematics* 12.2 (2001).

[12] StatSoft Inc. *Electronic Statistics Textbook, Chapter: Statistical Applications of Credit Scoring*. Feb. 2017. URL: http://www.statsoft.com/Textbook/Credit-Scoring.

[13] Stefan Lessmann et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". In: *European Journal of Operational Research* 247.1 (2015), pp. 124–136.

[14] Xian Liu. *Survival analysis: models and applications*. John Wiley & Sons, 2012.

[15] The Royal Bank of Scotland Group. *Annual Report and Accounts 2016*. Feb. 2017. URL: http://www.investors.rbs.com/~/media/Files/R/RBS-IR/results-center/annual-report-2016.pdf.

[16] Lyn C Thomas. "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers". In: *International journal of forecasting* 16.2 (2000), pp. 149–172.

[17] Harry Zhang. "The optimality of naive Bayes". In: *AA* 1.2 (2004), p. 3.