Table of Contents

1. Academic Paper Review

The existing literature on data mining is extensive and focuses on various aspects within the process.

Lahmiri (2016) provides us an overview of how to construct and organize our work in the context of credit risk assessment. In his study of financial risk classification, Lahmiri (2016) compared several predictive models that combine feature selection techniques with data mining classifiers in terms of accuracy, sensitivity and specificity statistics. His idea of introducing models to make comparison reminds us of using different machine learning algorithms and picking the best-performing one as our final product. As suggested by Lahmiri (2016), SVM showed the best accuracy in his study and is also an appropriate classifier in real applications. Therefore, we deliberately include SVM in our alternative models to check if it can also provide outstanding accuracy on our dataset.

We have put spotlight on feature selection during review of academic papers. A considerable amount of literature has been published on feature selection field. From theoretical and practical insights provided by these studies, we have received instructive guidance in efficient data pre-processing.

Yu & Liu (2004) introduces a new feature selection framework. The framework decouples relevance analysis and redundancy analysis in finding optimal feature subset. In their review of existing feature selection methods, Yu & Liu (2004) pointed out that major focus was on relevant predictors, which was not sufficient in feature selection of especially high-dimensional data. Although the data we are dealing with is not high-dimensional, we have repeated measures in several different periods. Their study has raised our concern to introducing redundancy analysis into our project. On top of this, we are also inspired by their approach of efficient elimination of redundant features via explicitly handling feature redundancy. Following the outline of their novel framework, we have split our selection process into two stages: first, a variance-based relevance analysis, and second, a correlation-based redundancy analysis.

For relevance analysis, we have tried to find various feature selection methods to test their efficiency on our data set. Chandrashekar & Sahin (2014) provides a thorough overview of existing methods. We have learnt that these means can be broadly grouped into three main types: Filter, Wrapper and Embedded. Filter methods use variable ranking techniques to filter out irrelevant features. Wrapper methods use learning algorithm as a black box to evaluate the performance of variable subset. Embedded methods incorporate variable selection as part of the machine learning process. Under each main type, Chandrashekar & Sahin (2014) also discussed some of specific measures that can be utilized in practice. We find this literature helpful in leading us to a clear list of methods that can be used, as well as an understanding of basic concept and characteristic of each method. While we are most familiar with entropy-based information gain, which belongs to filter methods, we also want to include perspectives from other method to enhance features screening. Random forest importance from embedded methods has caught our attention for its exceptional adaptivity to random forest model.

Rodriguez-Galiano *et al.* (2018) draw our attention to a comparison of different feature selection approaches in the context of application. Their study suggests certain details worthy of attention during the implementation of feature selection. Especially, we learn that information gain is biased in favor of features with more values. In place of information gain, the gain ratio ranker normalizes the gained entropy values with their corresponding entropy (Rodriguez-Galiano *et al.*, 2018). Therefore, we plan to choose gain ratio instead of information gain as filter criteria for our feature selection process. Finally, we have a more detailed look on random forest model. Huang & Boutros (2016) examined the parameter-sensitivity of RFs in computational genomic studies. In their analysis of random forest parameterization, Huang & Boutros (2016) tested the effects of parameter selection on classification performance on two biological datasets with distinct p/n ratios. The result turned out that model with default parameters performed well generally but is not the optimal choice. This result surprises us by questioning the reliability of default parameters. We have learned that the optimal m_{try} value can be determined systematically by certain function, while tuning n_{tree} and *sampsize* to improve the model performance is still open for our experiments.

2. Data Mining Process

2.1 Business Understanding

Universal Plus, our potential client, loans money to their clients to be paid in a fixed term. Due to the absence of credit level assessing procedure, the firm has been exposed to high risk of default. Therefore, our primary goal is to help Universal Plus establish a credit risk management system. The system will provide clear guidance in targeting premium clients and anticipating potential defaults, and hence contribute to overall profit improvement.

We have a dataset provided by Universal Plus that contains 31,375 instances of historic loan applications. We use R as the software to process data and build up default prediction models. Computer crash is a potential risk that could delay completion of this project. To counter this, we back up our project file on daily basis in case of progress loss after a sudden hardware or software crash.

Our deliverables have three components: a default prediction model in R, a report elaborating the modeling process and an overview of our outcomes through video presentation. Success of the default prediction model is defined as model accuracy larger than 80% as well as recall rate larger than 50%.

2.2 Data Understanding

Universal Plus have provided us a dataset of historic loan applications, including personal characteristics, recent transaction record and credit record for each client. The full dataset has 31,375 observations in 39 different variables.

Our primary focus is on CLASS variable, which indicates default and non-default cases. Among all the client records that can be seen, 6,966 of them go into default category

and the rest 24,409 are non-default. As positive (default) records take up only 22%, the dataset is skewed to negative class.

According to the data dictionary, there are 23 factorial attributes and 13 numerical attributes. We plotted each numerical variable to spot any potential outliers. The graphs show that all these numerical variables related to credit limit, bill statement amount and repayment amount share a similar distribution pattern. While most values are dense around a low medium value, a long tail and dispersed outliers can be seen along the x-axis. By checking instances with extreme values carefully, we decided to keep all these records as they represent certain situations that could happen in daily business.

2.3 Data Preparation

Basic data processing has been carried out as follows:

- 1,339 duplicate observations were dropped.
- 210 observations containing missing values were dropped.
- Several variables were adjusted from numerical type to factorial type according to its meaning.

Upon closer inspection of data, we were surprised to find that for a considerable number of instances, their bill statement amount exceeded the credit limit amount. We supposed this abnormal exceed might have come from temporary adjustment of credit limit. We decided to keep all these records and gain some extra information from them. So, we derived a new attribute named “OVERDRAFT” to indicate if any bill statements went beyond limit or not. After data-preprocessing above, we split the data into training and test set. Then, the training set were balanced to mitigate data skew.

Once the training data set was ready for modeling, it was first necessary to select informative predictors and exclude attributes of no use. Following the feature selection framework from Yu & Liu (2004), we split the process into relevance analysis and redundancy analysis.

Due to the broad sample size of our dataset, filter and embedded methods are generally more efficient in computation time than wrapper method. Therefore, for relevance analysis criteria, we chose random forest importance for random forest model exclusively, and gain ratio for all other models. At this stage, we kept features that have mean decrease in Gini coefficient larger than 200 in random forest model (22) and features that have positive gain ratio values for other models (32).

Once relevant features have been determined, the redundancy analysis is started by computing correlation between every pair of features. As the relevant features we selected are a mixture of both nominal and numerical ones, the strength of correlation is computed for nominal vs nominal with a bias corrected Cramer's V, numeric vs numeric with Spearman correlation, and nominal vs numeric with ANOVA. Using association score 0.75 as an indicator of high correlation, it was possible to identify the shared variances among bill statement amount in each period. Integrating their rank in relevance, less-informative bill statement attributes were then removed from training set.

2.4 Modeling

Prior to building models, we selected five modeling techniques to be used and compared: the random forest (RF), the decision tree (Tree), the logistic regression model (LR), the support vector machine (SVM) and the gradient boosting machine (GBM). A common advantage of these model is that their output can be interpreted easily. Our client will then have fewer barriers to understand and use the model.

Upon completion of feature selection, five models were built on the training set. For random forest model, we tuned three key parameters (n_{tree}, m_{try} and *sampsize*) to optimize model performance. Several sets of parameters were created and evaluated. The default value for n_{tree} (n = 500), tuned m_{try} (n = 2) and *sampsize* (n = 5000) were observed as the optimal parameterization.

The rest of models (Tree, LR, SVM and GBM) were built on features selected by gain ratio and redundancy analysis. The prediction capacity of each model was then validated on test set and presented by a confusion matrix.

Table 1
Summarised results of 5 modeling techiniques. MLA: Machine Learning Algorithm; RF: Random Forest; Tree: Decision Tree; LR: Logistic Regression Model; SVM:Support Vector Machine; GBM: Gradient Boosting.

MLA	Accuracy	Recall	Features selected
RF	0.8007	0.5376	PY1, BILL1, LIMIT, PYAMT1, BILL2, AGE, PYAMT2, PY2, PYAMT3, PYAMT6, PYAMT4, PYAMT5, PY3, PY4, PY5, YEARSINADD, PY6, EDUCATION
Tree	0.7801	0.5782	PY1, PY2, PY3, PY4, PY5, PY6, PYAMT1, LIMIT, PYAMT2, PYAMT3, OVERDRAFT, PYAMT5, PYAMT4, PYAMT6, EDUCATION, BILL5, GENDER, AGE, MARRIAGE, EMPLOYMENT, RSTATUS, OTH_ACCOUNT, FREQTRANSACTION, DEPENDENT, NEW_CSTM, SECONDHOME, SATISFACTION, CAR, PHONE
LR	0.8007	0.5138	PY1, PY2, PY3, PY4, PY5, PY6, PYAMT1, LIMIT, PYAMT2, PYAMT3, OVERDRAFT, PYAMT5, PYAMT4, PYAMT6, EDUCATION, BILL5, GENDER, AGE, MARRIAGE, EMPLOYMENT, RSTATUS, OTH_ACCOUNT, FREQTRANSACTION, DEPENDENT, NEW_CSTM, SECONDHOME, SATISFACTION, CAR, PHONE
SVM	0.7874	0.5115	PY1, PY2, PY3, PY4, PY5, PY6, PYAMT1, LIMIT, PYAMT2, PYAMT3, OVERDRAFT, PYAMT5, PYAMT4, PYAMT6, EDUCATION, BILL5, GENDER, AGE, MARRIAGE, EMPLOYMENT, RSTATUS, OTH_ACCOUNT, FREQTRANSACTION, DEPENDENT, NEW_CSTM, SECONDHOME, SATISFACTION, CAR, PHONE
GBM	0.7866	0.5560	PY1, PY2, PY3, PY4, PY5, PY6, PYAMT1, LIMIT, PYAMT2, PYAMT3, OVERDRAFT, PYAMT5, PYAMT4, PYAMT6, EDUCATION, BILL5, GENDER, AGE, MARRIAGE, EMPLOYMENT, RSTATUS, OTH_ACCOUNT, FREQTRANSACTION, DEPENDENT, NEW_CSTM, SECONDHOME, SATISFACTION, CAR, PHONE

2.5 Evaluation

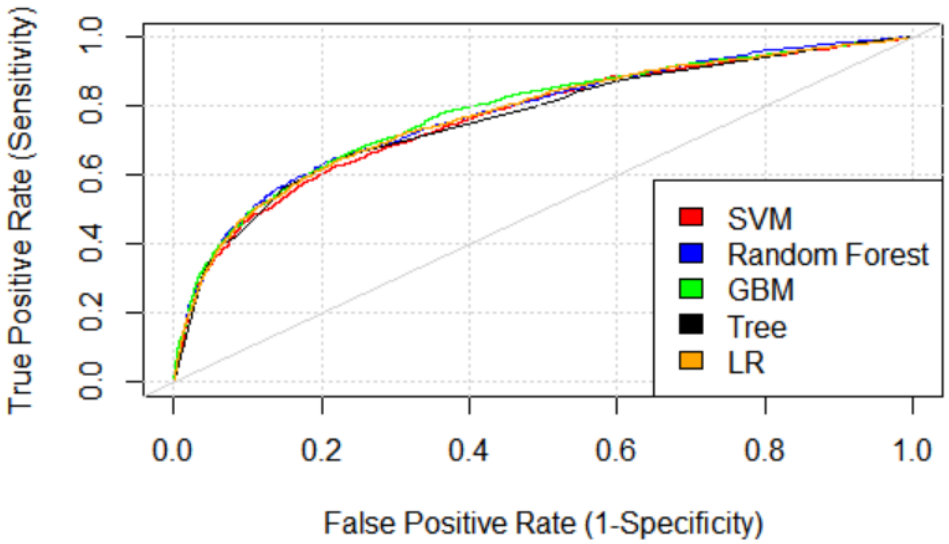

Fig. 1. ROC curves of the five models: SVM: Support Vector Machine; GBM: Gradient Boosting; Tree: Decision Tree; LR: Logistic Regression Model.

As our primary goal is to identify potential default cases for Universal Plus to avoid credit risk proactively, the ideal model should perform well in spotting positive instances. So, the evaluation of models is first based on recall rate (Table 1). As regards the recall rate, the decision tree gave the best prediction on true positive cases at the expense of false alarm. A bias on positive side will lead to reduced accuracy. On one hand, we hope the model to detect as many default instances as possible. On the other hand, it is also our aim to minimizing the side effect of losing creditworthy clients. Therefore, decision tree was eliminated from our options.

Searching downwards by recall rate, GBM and random forest appeared as a pair of complementary options. While GBM could provide the second highest recall rate, its performance in accuracy was poorer than random forest. An AOC graph (Fig. 1) and AUC figures (0.7795 for GBM and 0.775 for RF) also indicated that these two models have comparable predictive capability overall. Taking application scenario into account, the model will generally encounter far more non-default cases than default cases. Although the percentage differences in recall rate and in accuracy of two models are similar, the larger amount of non-default instances will amplify GBM's mistake in raising false alarm and causing unnecessary client loss. Therefore, we decided to choose random forest as our final product.

2.6 Deployment

By importing new loan application data to the model, the prediction result will enable Universal Plus to make more considerate decision in future client screening. Specifically, if the prediction result turns out as non-default, we think Universal Plus can accept the application since 87% of cases the client will not default indeed. If there is default warning suggested by the model, however, the firm should at least conduct more detailed credit check before lend money to that client.

To ensure that the prediction model will be used properly on an ongoing basis, we plan to conduct model performance check at the start of every quarter. This check will be based on client records from past quarters. Any necessary adjustment and update will be provided before the model being used for the upcoming period.

In conclusion, this project was undertaken to design a credit management system that enables proactive credit control of Universal Plus. Using the dataset provided by the firm, we identified informative predictors and built up a random forest model with 53.76% default detection rate.

References

[1] Lahmiri, S. (2016) Features selection, data mining and finacial risk classification: a comparative study. *Intelligent Systems in Accounting, Finance and Management*, 23(4)：265-275.

[2] Yu, L. & Liu, H. (2004) Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5: 1205-1224.

[3] Chandrashekar, G. & Sahin, F. (2014) A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1): 16-28.

[4] Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M. & Mendes, M. P. (2018) Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of the Total Environment*, 624: 661-672.

[5] Huang, B.F. & Boutros, P.C. (2016) The parameter sensitivity of random forests. *BMC Bioinformatics*, 17: 331. Available from: https://doi.org/10.1186/s12859-016-1228-x (Accessed 3 December 2021)