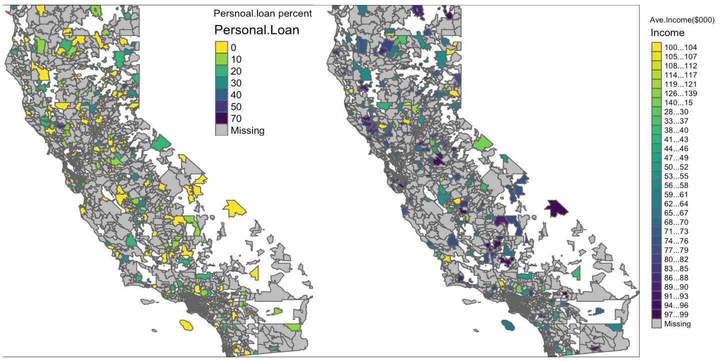
Classification Report

Student ID: tbdr69 / Name: Minwoo Kim

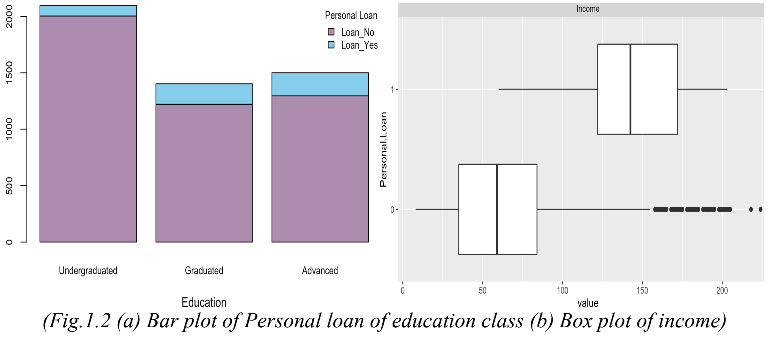
Part I: Executive Summary

For upselling in a bank institution in California district in the USA, further analysis led to the cause of more customers in certain conditions having a loan. Since the data provided these results with highly correlated elements about a loan, the bank needs to know what customers are more likely to loan and establish targeting crucial elements for marketing campaigns for upselling.
So, the objective is to find the **most critical factor** that promotes marketing campaigns most and provide guidance to locate customers that are willing to loan.



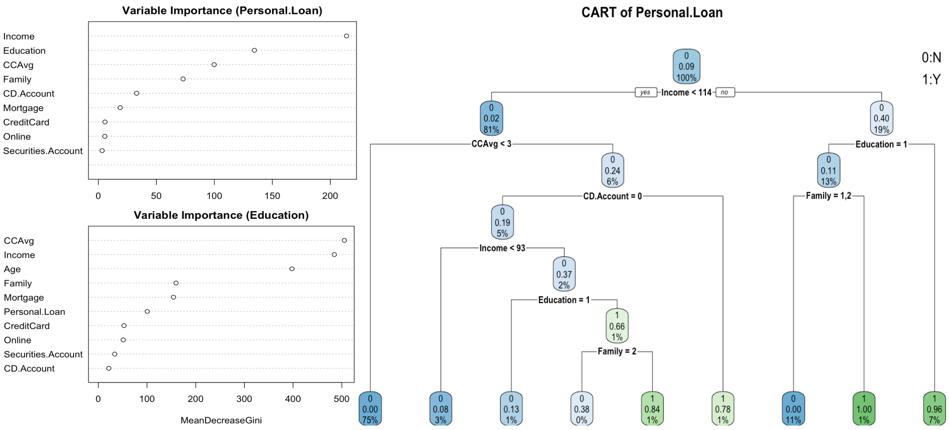
(Fig.1.1 (a) Heatmap of Personal loan (b) Income in California area)

Based on the zip code data, we can see from the personal loan Fig.1.1(a) related to the district, and it also shows the **average education level** and **income** by zip code Fig1.1(b). This indicates that the **region** may play an **important role** in making loan marketing campaigns.



(Fig.1.2 (a) Bar plot of Personal loan of education class (b) Box plot of income)

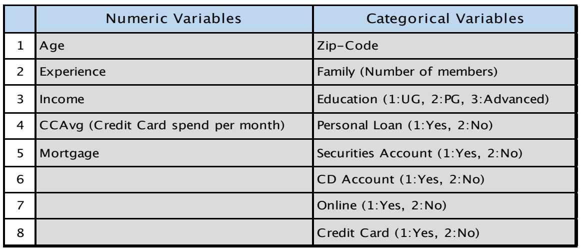
Fig1.2(a) shows that customers who loan from the bank have **high education levels**. The reason may be that they need more money than an expense from education compared with other education levels.
From Fig1.2(b), we can see that customers who are likely to be **high income** tend to get more loans in the bank, and some of the high income did not have a loan which represents outliers. Probably because high incomes are easier payback to the bank and some high income does not need a loan.



(Fig.1.3 (a) Important variables in Personal loan and Education (b)Tree classification result)

The importance analysis of our model is shown in Fig1.3(a). As we can see in the figure, **income** is at the top of the importance rank in personal loan based analysis, meaning that income greatly influences whether the customer loan or not.
Moreover, the second important part was **education**. Furthermore, **education is highly related to income**. from education based analysis, From analyses, **Personal Loan, Education, and Income have a major part** to consider upselling marketing in the bank.
And in Fig.1.3(b) personal **loans have about 20%** of California people. And we can know the threshold for **income is 114,000$**. It is about 19% of loaned people if the education level is over undergraduate, about 7% of loan people. So, it is the highest group of loaned people. On the contrary, average credit card spending per month **under 3000$ is 75% of not** being loaned from banks.
As a result, the most premium target for upselling would be **high income** and **over undergraduate** education level.
If the bank wants to expand the customers, they need to use marketing to **credit card users** who **spend under 3000$** in a month.

Part II: Technical summary



	Numeric Variables	Categorical Variables
1	Age	Zip-Code
2	Experience	Family (Number of members)
3	Income	Education (1:UG, 2:PG, 3:Advanced)
4	CCAvg (Credit Card spend per month)	Personal Loan (1:Yes, 2:No)
5	Mortgage	Securities Account (1:Yes, 2:No)
6		CD Account (1:Yes, 2:No)
7		Online (1:Yes, 2:No)
8		Credit Card (1:Yes, 2:No)

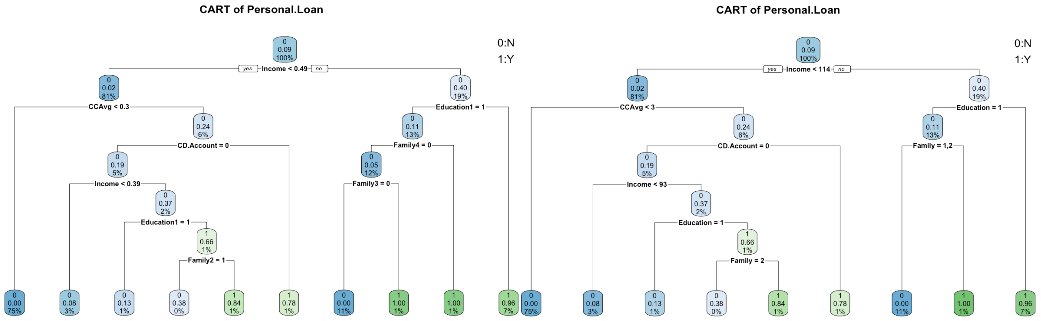
(Table.2.1 Datatypes)

- Initial data summary

This dataset is customers loan data (with personal information) and consists of 5000 customer data on the bank of California. This dataset has 1 classification response in which the customer has a personal loan (1: yes or 0: no) or not. And it contains **13 variables**, including **5 numeric variables and 8 categorical variables** in Table.2.1

- Strategies with missingness

There are some unknown data in numeric variables, which is the **experience variable**, which has **negative values**, but an age variable is highly correlated with experience. So, the experience variable was **deleted**. Moreover, "*incorrect*" values in categorical variables in the Zip-Code variable are **eliminated** as missing data only for map heatmap classification.



(Fig.2.1 (a) Standardised Tree result (Left) (b) Original Tree result (Right)

- Data processing or feature transformation

For categorical variables, on this classification model changed the **data type** in a data frame. Therefore, programs recognise these variables as categories as a character.

Secondly, the data was standardisation for numerical variables. And it compared with not standardised results in the CART model in Fig.2.1(a) and (b). But it has not that different results. So used the **original version** for convenience.

- Train/Test/Validate

The data was split into two parts: **training set (70%)** and **testing set (30%)** with random seed(1) for reproducibility. A training set is used for initial model fitting, and a testing set is for model assessment.

- Model design

On this classification built three models: **Random Forest, CART**(Classification and Regression Tree) **and Bagging**. Beyond this, there was LDA (Linear Discriminant Analysis) to compare with the previous three models. Nevertheless, it has low accuracy than the other three models, and for the graph, it needs three groups, but the Personal Loan variable has only two groups. Therefore, it is used only for additional results in this report. To compare three models, three methods as the evaluation metric. After training, fitted models are assessed and compared through validating set. Finally, this report chose **accuracy, ROC**, and **AUC** as assessment metrics.

- Model selection

The bank refers to customers who are **willing to have a loan** as valuable customers because of profit increase.

True-positive represents the predicted valuable customers are the valuable ones. Conversely, False-positive indicates the resource wasted for acquiring predicted valuable customers who will not loan. Hence, accuracy was the most important part of model selection.

The model was selected by **helpfulness to the bank's upselling** strategy, and the selected model has assessed over 99% in the accuracy and AUC for making upselling strategies for banks. In the case of Random Forest, it will run several decision trees and vote for each tree. It is useful for finding which variable is **most influential for target** classification. So, it was found that **income** is most influential for the personal loan, and the second was education.

From the second important variable, rerun the random forest, and it is also big relation between income and CCAvg. Therefore, income and **education** have been decided as powerful variables for personal loans (CCAvg is highly related to education but low related to a personal loan).

In the case of CART, it is selected because it is easy to understand which group have certain decisions. in the **visualisation** parts, CART is an excellent way to use it.

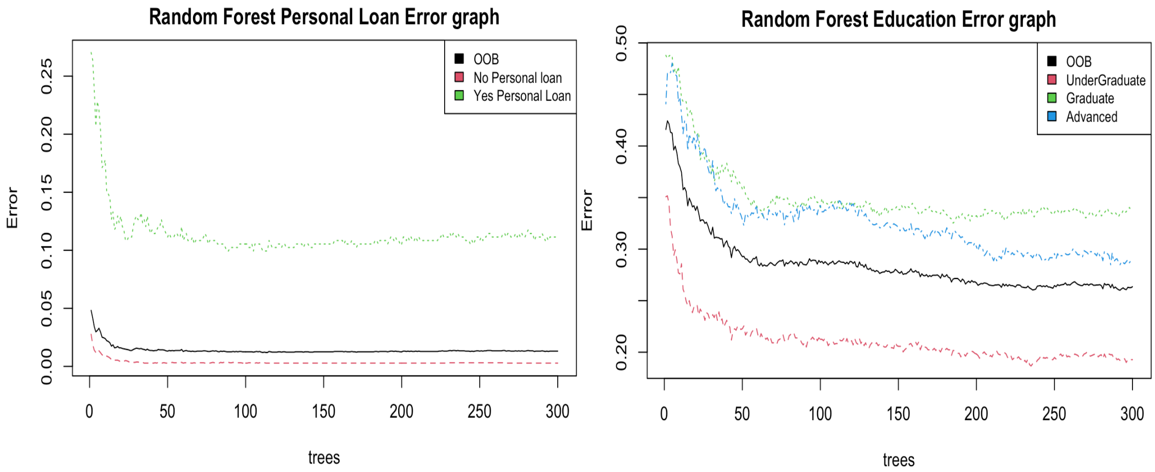
However, the CART is slightly less accurate than random forest, so it needs to boost the result. So, the bagging makes training data into several data and does the classification parallelly. With this bagging, cross-validation was also performed, which makes bagging more accurate.

This report compared ROC and AUC with these models for measuring how well the model performs in classification. AUC closer to 1 indicates better performance.
As we can see in the figure follows, the random forest has the highest accuracy and AUC. So, this report selects **Random Forest** as the most **suitable model** for this classification.

- Hyperparameter tuning

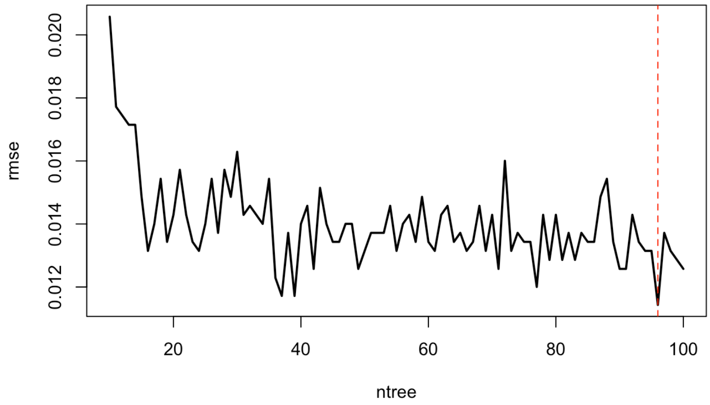
There are some hyperparameters that we need to tune for random forest model and the main hyperparameters list as below: Number of trees to grow in the forest ntrees Number of variables randomly sampled as candidates at each split mtry Size(s) of sample to draw sampsize Minimum size of terminal nodes node size Maximum number of terminal nodes trees in the forest can have max nodes Among them, **ntrees** and **mtry** have the largest influence on predictive accuracy, and I focused on these two parameters.
Usually, we need sufficiently large ntrees to ensure that the error rate stabilises at a certain level. In this case, I chose a set of values 500 trees to start with. but it was enough to use **300 in ntree** value.
For mtry, this report tried to use 7 or 10 mtry, but it has a high error rate than the **default value**. So set the mtry as default.
In the CART, use the control hyperparameter in rpart. Control this parameter is to save computing time by pruning off splits that are not worthwhile. This parameter sets the **complexity parameter** as 0.0001 to further prune the **cp value** (with a more parsimonious tree), minimising the xerror (cross-validation error).
In the Bagging, for increasing accuracy, using **cross-validation** for traincontrol function. Furthermore, using the **10 folds** to cross-validate and find the least root mean square error. Also, it used the importance option to find the most valuable variable.

- Model performance and interpretation



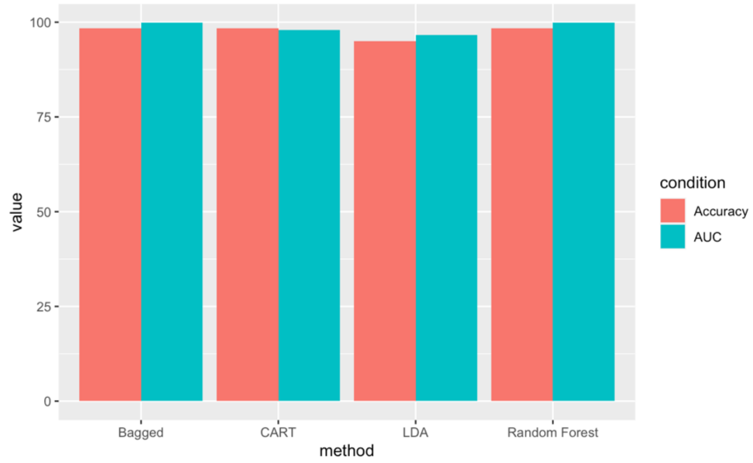
(Fig.2.2 (a) Random Forest Error graph based on Personal Loan (b) Random Forest Error graph based on Education)

In the Random Forest model, errors like in Fig.2.2(a). In the figure, a personal loan based analysis is relatively **stable** with Random Forest. It was converged after **50 trees**. OOB(Out of Bag) error (0.025) was also small than Fig.2.2(b). In Fig.2.2(b), it was an education based analyse. It also converged after **200 trees**, and OOB was 0.25. This was **10 times higher** than personal loan based analysis. OOB Error is the number of wrongly classifying the OOB Sample. That means data has not been used while training the model in any way, so Fig2.2(a) was not any **leakage of data** and henceforth ensured a better predictive model than Fig.2.2(b).



(Fig.2.3 RMSE by number of tree in Bagging)

In the Bagging, RMSE(Root Mean Square Error) have turbulence during the number of trees. During 100 iterations, bagging finds the lowest RMSE value. It is 0.011, and if the Bagging sets the tree as **96 trees**, RMSE would be the lowest.

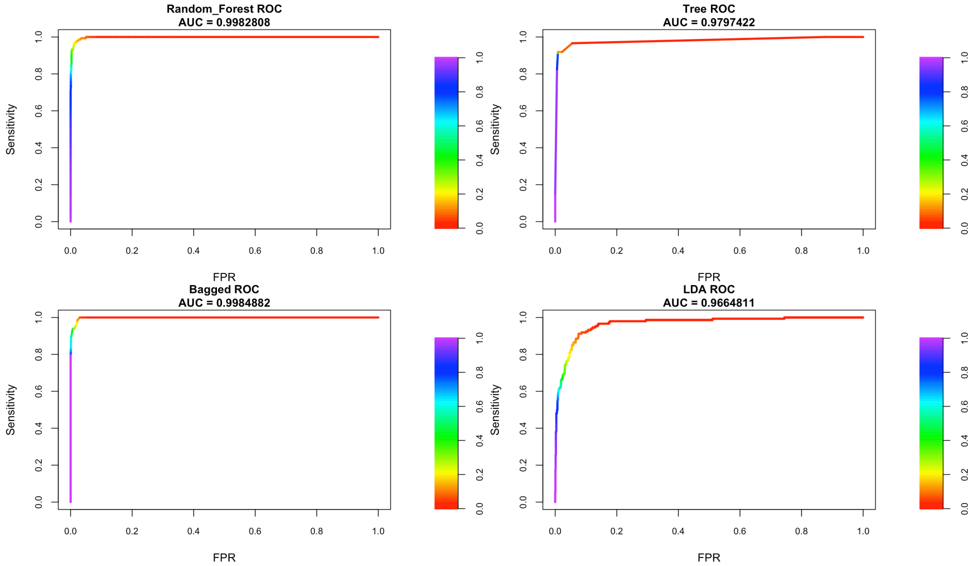


(Fig.2.4 Bar plot of Accuracy and AUC results by Models)

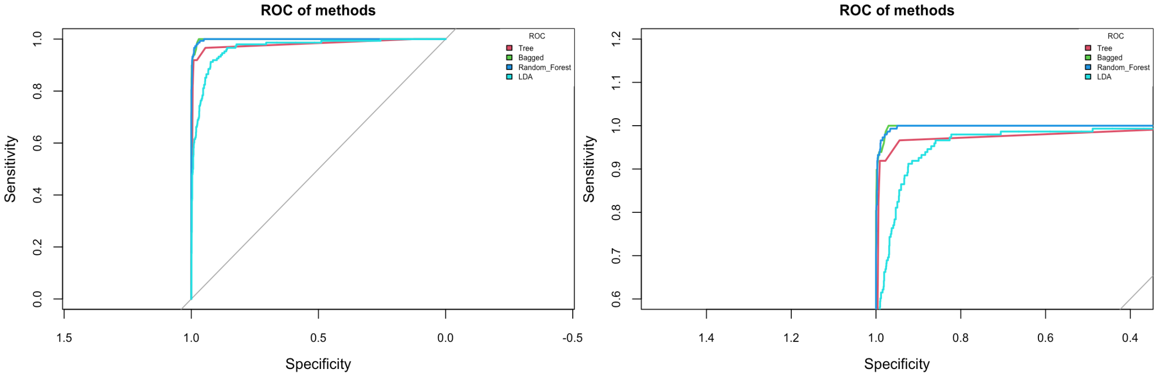
Bagged		CART		LDA		Random Forest	
Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
98.33	99.82808	98.33	97.97422	94.93	96.6481	98.67	99.84882

(Table.2.2 Value of each model results)

Above the result, in Fig.2.4 and Table.2.2, the actual value of accuracy and AUC (Area Under Curve) can be found. Referencing the table in **AUC**, the model fitted well by the following sequence, Random Forest(99.85) > Bagging(99.83) > CART(97.98) > LDA(96.65) performs well in predicting. In Table.2.2, **accuracy** between prediction and test data also follows AUC sequence like Random Forest(98.67) > Bagging(98.33) = CART(98.33) > LDA(94.93) In here, Bagging and CART have the same accuracy in test data. It may indicate that CART was well-fitting in this test set, but the Bagging equalizes influence, so Bagging could reasonably fit in other test sets. Bagging also reduces the variance. However, Bagging is not always perfect.



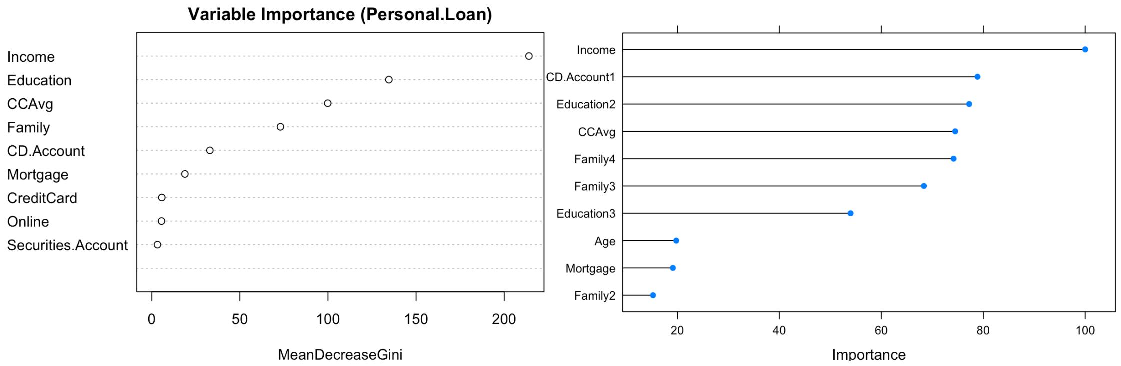
(Fig.2.5 ROC Curve with AUC value by Models)



(Fig.2.6 ROC Curve with all models and zoom in the figure)

For ROC and AUC, as it shows, the AUC value is over 0.99 in Random Forest and Bagging, which is very **close to 1**, indicating that this model **performs well** in classification.

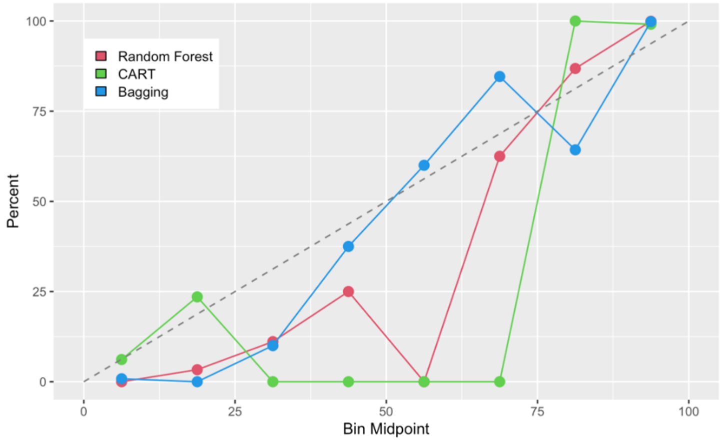
In Fig.2.5, all models have excellent conditions in the ROC curve. Some models have nearly rectangle shapes, which means that models are perfect for this classification. From Fig.2.6, much information could get about models. In CART model has almost near the shape with Random Forest and Bagging. After boosting with **Bagging**, it **improves** the **CART** model.
In addition, Bagging and Random Forests have **almost the same** shape in ROC Curve. Because Bagging and Random Forest use the similar algorithm. The fundamental difference is that in Random Forests, only a **subset of features** is selected randomly out of the total. Unlike Bagging, the best split feature from the **subset is used to split** each node in trees, where all features are considered for splitting a node.



(Fig.2.7 (a) Importance variables in Random Forest (b) Importance variables in Bagging)

At last, a random forest model can generate the **importance rank** for each variable. As shown in Fig.2.7, the importance value of **income** is much higher than all the other variables, and the second was an **education** in Fig2.7(a). In Fig2.7(b), Bagging indicates that having a certificate of deposit with the bank is the second important variable. However, it changes every Bagging tree. On the other hand, Education (PG) was rather high related to personal loans in the Bagging model. Therefore**, income** and **education** are the most **valuable data** for a personal loan important role for an **upselling strategy** which got from Random Forest and Bagging.
For this, variables and Map data could enforce the establishing upselling strategies following the Part 1 report.

- Improvement for next Classification



(Fig.2.6 Calibration Curve with all models)

Calibration reflects how well the predicted class **probabilities** match the 'true' probabilities according to the underlying distribution of the data. Consequently, the properties of a model algorithm itself **do not universally** determine how well- or poorly calibrated the results will be.
From bank loan data, some of the variables were partially covariates, or some of the variables may not affect the aimed classification variable.
Hence being well-calibrated or poorly calibrated **depends on the problem and data quality** and is not a universal property. For improvement of calibration problems or well Classification, in this problem, using Logistic Regression produce a reasonable estimate of the probability. Furthermore, if the Random Forest and Logistic Regression combined, it could try to compute a **weighted** average of the **variable importance** (maybe after normalizing each variable importance vector to unit length) for various values and the averaging weight and then pick up the value that yields the best **cross-validated score** for the final model.

Part II: Reproducible code

Link 1: (Google Drive Download)
https://drive.google.com/file/d/1gvHnEn4Zd3JUcQ0dcnbjXbokNUAdhmov/view?usp=sharing

Link 2: (Github)
https://github.com/MinwooKim1990/MinwooKim1990/tree/Classification