INTRODUCTION

Background
Credit risk is one of the eight risks that banks must consider. It is important to make a measurable, documented, and developable credit risk system. Logistic regression, discriminant analysis, and artificial neural network are some methods that are used in credit risk model. They are useful to predict whether a new applicant will become a good or bad debtor if he or she receives a loan.

Multicollinearity is a common problem in credit risk modeling. Usually, the solution for this problem is using variable selection method (forward, backward, and stepwise). But this solution may cause missing information about the response variable if the deleted predictor variable is an important one. Ridge regression is another statistical procedure for dealing with the problem of multicollinearity (Ravinshanker & Dey 2001). With logistic ridge regression, the multicollinearity is expected to be handled without deleting any variables and there will be no missing information from the data that has been collected.

Bank of Indonesia noted that the growth of credit of national banks in January 2010 was 10%. Until the end of August 2010, the credit of banking industry grew and reached 20.3% (Purnomo 2010a, 2010b). This may conduce on a greater risk that has not been faced by banks before. Hence, it is important to build a more accurate credit scoring model to decide whether a new applicant is credible enough to get a loan.

Objectives
The objectives of this research are:
1. To build a credit risk model using logistic regression with variable selection and logistic ridge regression.
2. To determine the optimal probability cutpoint.
3. To compare the classification rate and the c statistic of logistic regression with variable selection and logistic ridge regression.

LITERATURE REVIEW

Credit Risk Model
Banks loan to individuals, first by asking to fill out a loan application. The customer is asked to submit several documents that the bank needs in order to evaluate the loan request. There are six aspects of the loan application to determine whether a new applicant is creditworthy or not. The Six Basic Cs of Lending are namely character, capacity, cash, collateral, condition, and control. Character is the data about the personality. Capacity is the capacity to borrow money. Cash is related to the borrower income and balance in saving account. Collateral is the adequacy of the borrower to provide adequate support for the loan. Age and degree of specialization of the borrower's assets are the example of collateral. Condition is the prospect of business associated with economics conditions. Correctly prepared loan document is the example of control.

The basic theory of credit scoring is that the bank can identify the financial, economic, and motivational factors that separate the good debtors from the bad ones by observing a large group of people who have borrowed in the past. Credit scoring systems are usually based on discriminant models or related techniques such as logit or probit models or neural networks. If the applicant’s score exceeds a critical cutpoint level, he or she is more likely to be approved for credit. Among the most important variables used in evaluating consumer’s loan are age, marital status, number of dependents, home ownership, telephone ownership, type of occupation, and length of employment in a current job.

The Cramer Statistic
The chi-square test of independence is used to conclude whether there is an association between two categorical variables. When the number of rows and columns of the contingency table are unequal, Cramer coefficient is the measure of the strength of this association. The value is between 0 and 1. The Cramer coefficient is defined as:

\[ C = \frac{X^2}{n(t-1)} \]

Where \( X^2 \) is the chi-square statistic, \( n \) is the total sample size, and \( t \) is either the number of
rows or the number of columns in the contingency table, whichever is smaller.

**Logistic Regression**

Let the conditional probability that the outcome is present be denoted by \( P(Y=1|x) = \pi(x) \). The logit of the multiple logistic regression is given by the equation

\[
g(x) = \logit(\pi(x)) = \log \left( \frac{\pi(x)}{1-\pi(x)} \right)
\]

where \( \pi(x) \) is the likelihood function with \( p \) covariates. Under the null hypothesis, the likelihood without covariates, and

\[
L_0 = \log \left( \frac{\pi(x)}{1-\pi(x)} \right)
\]

in which case the logistic regression model is

\[
\pi(x) = \frac{e^{\theta(x)}}{1 + e^{\theta(x)}}
\]

When the type of independent variable is categorical, dummy variable is needed. In general, if a categorical variable (nominal or ordinal scale) has \( k \) values, then \( k-1 \) design variables will be needed. Thus, the logit for a model with \( p \) variables and the \( j^{th} \) variable being categorical would be

\[
g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{i=1}^{k-1} \beta_j D_i + \beta_p x_p
\]

Maximum likelihood estimators to logit model are obtained by maximizing \( \beta \) of the likelihood function

\[
l(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]
\]

After getting the model, we begin the process of model assessment. The significance of the covariates could be assessed by G test statistic and Wald test. G test statistic is a likelihood ratio test and measures the significance of the parameters on the overall model. Hypothesis of G test statistic:

\[
H_0: \beta_i = \beta_2 = \cdots = \beta_p = 0
\]

\[
H_1: \text{at least one } \beta_i \neq 0, \ i = 1, 2, \ldots, p
\]

G-test Statistic could be formulated as

\[
G = -2 \ln \left( \frac{L_0}{L_p} \right)
\]

where \( L_0 \) is Likelihood without covariates, and \( L_p \) is Likelihood with \( p \) covariates. Under the null hypothesis, the distribution of \( G \) is chi-square \( \chi^2 \) with \( p \) degrees of freedom.

If the null hypothesis is rejected and conclude that at least one and perhaps all \( p \) coefficients are different from zero, the Wald test could be used to assess the significance of each covariate.

\[
H_0: \beta_i = 0
\]

\[
H_1: \beta_i \neq 0 \text{ where } i = 1, 2, \ldots, p
\]

Under the null hypothesis, \( W \) statistic will follow a standard normal distribution (Hosmer & Lemeshow 2000).

Coefficient interpretation in logistic regression is by using the odds ratio that indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group. The odds ratio defined as \( OR = \exp(\beta) \). For numeric variable, the odds ratio indicates that for every increase of one measurement of the predictor, the risk of the outcome increases \( \beta \) times.

Multicollinearity can cause unstable estimates and inaccurate variances which affects hypothesis test (Hoerl & Kennard 1980, in Shen & Gao 2002). In regression, there are some approaches to handle multicollinearity, which are variable selection method (forward, backward, and stepwise) and using ridge regression. Forward selection adds terms sequentially until further additions do not improve the model. Backward elimination begins with a complex model and sequentially removes terms. Stepwise procedure starts off by choosing the equation containing the most important variable and then attempts to build up with subsequent additions of variable one at a time as long as these additions are worthwhile.

**Logistic Ridge Regression**

Unstable parameter estimates occur when the number of covariates is relatively large or when the covariates are correlated. An alternative procedure to obtain more stable estimates is to specify a restriction on the parameters. Consider the maximization of the log-likelihood function with a penalty on the norm of \( \beta \):

\[
l^A(\beta) = l(\beta) - \lambda \|\beta\|^2
\]

where \( \|\beta\| = (\sum \beta_i^2)^{1/2} \), the norm of the parameter vector \( \beta \). The ridge parameter \( \lambda \) controls the amount of shrinkage of the norm of \( \beta \). When \( \lambda = 0 \) the solution will be the ordinary MLE. For a good choice of \( \lambda \), the estimate \( \hat{\beta}^\lambda \) is expected to be an average closer to the real value of \( \beta \) than the ordinary MLE, i.e. \( \text{MSE}(\hat{\beta}^\lambda) < \text{MSE}(\beta) \) (Cessie & Houwelingen 1990).

The estimate parameter of logistic ridge regression is calculated in the following ways:
1. Fit the logistic regression model using maximum likelihood, leading to the estimate of $\hat{\beta}$. Construct standardized coefficients by defining

$$\hat{\beta}_j^* = \frac{\hat{\beta}_j}{s_j} \quad j=1,2,...,p$$

where $s_j$ is the standard deviation of $\beta$ in the training data for the $j$th predictor.

2. Construct the Pearson $X^2$ statistic

$$X^2 = \sum_{k=1}^{g} \frac{(y_k - m_k \hat{n}_k)^2}{m_k \hat{n}_k (1 - \hat{n}_k)}$$

where 
- $g$ = the number of covariate patterns
- $m_k$ = the number of subjects with $x=x_k$
- $y_k$ = the number of positive responses ($y=1$) among the $m_k$ subjects
- $\hat{n}_k$ = probability that the outcome is present in $x=x_k$

This is a measure of the difference between the observed and the fitted values.

3. Define the ridge parameter ($\lambda$)

$$\lambda = \left( \frac{p}{N - p - 1} \right) \sum_{j=1}^{p} (\hat{\beta}_j^*)^2$$

4. Let $X$ be the matrix of centered and scaled predictors, with

$$x_j = \frac{X_j - \bar{X}_j}{s_j}$$

Let $\Omega = Z'VZ$, where $V_N$ is $diag[\hat{\beta}_0(1 - \hat{\beta}_0)]$.

Let $\hat{\beta}^*$ equal $\beta$ with the intercept $\beta_0$ omitted. Then the ridge regression estimate equals $(\hat{\beta}_0^R, \hat{\beta}^R)$, where

$$\hat{\beta}_0^R = (\Omega + diag(2\lambda))^{-1} \Omega \hat{\beta}^*$$

and

$$\hat{\beta}_j^R = \hat{\beta}_j + \sum_{j=1}^{p} (\hat{\beta}_j^* - \hat{\beta}_j) X_j$$

Optimal Cutpoint

Optimal cutpoint for the purpose of classification can be obtained from the plot of sensitivity and specificity versus all possible cutoffs (Hosmer & Lemeshow 2000). The plot can be seen in Figure 1. The optimal cutpoint is not the only criteria for deciding whether a new applicant is acceptable or not to get a loan. Although the correct classification rate is high based on the optimal cutpoint, the number of false positive should be considered because the loss caused by this error is extremely large relative to the false negative. Each bank has its own criteria for making a decision. Explanation for these errors can be seen in the next session. In this research, the cutpoint score will just be attained from the plot of sensitivity and specificity versus all possible cutpoints.

![Figure 1: Plot of sensitivity and specificity versus all possible cutpoints](image1.png)

**Model Evaluation**

In model assessment, a classification table is most appropriate when classification is a stated goal of the analysis. Figure 1 is the classification table. It is a two way frequency table between actual data and the prediction. Correct classification rate (CCR) consists of percentage of true positive and true negative, while misclassification rate (MCR) consist of percentage of false positive and false negative.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>1</td>
<td>False Negative (FN)</td>
</tr>
</tbody>
</table>

![Figure 2: Classification Table](image2.png)

Sensitivity or true positive (TP) is the number of observation that have category 1 and was correctly predicted. Specificity or true negative (TN) or is the number of observation that have category 0 and was correctly predicted. False positive is the number of observation that have category 0 but predicted as category 1. False negative is the number of observation that have category 1 but predicted as category 0.

![Figure 3: ROC curve](image3.png)
Figure 3 shows ROC curve. It plots the probability of false positive (1-specificity) against true positive (sensitivity). The area under the ROC curve (AUR), which ranges from 0 to 1, provides measure of the model ability to discriminate between those subjects who experience the outcome of interest versus those who don’t. The measure of AUR is c-statistic.

\[ C = \frac{n_c + 0.5(t - n_c - n_d)}{t} \]

where

- \(n_c\) : the number of concordant
- \(n_d\) : the number of discordant
- \(t\) : the number of total pairs

As general rule:
- \(C = 0.5\) : no discrimination
- \(0.7 \leq C < 0.8\) : acceptable discrimination
- \(0.8 \leq C < 0.9\) : excellent discrimination
- \(C \geq 0.9\) : outstanding discrimination

(Hosmer & Lemeshow 2000).

**METHODOLOGY**

**Data Source**

The data used in this research was the German Credit data set which was available at http://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/. It contains observations on 1000 past credit applicants. Each applicant was rated as “good” (700 cases) or “bad” (300 cases). There were 17 variables used in this research after considering The Six Basic C of Lending which consist of 3 numeric variables, 6 ordinal variables, 7 nominal variables, and 1 binary variable. Description of the variables can be seen in Appendix 1.

**Method**

Procedures used in this research were:

1. Divide the data into training data (740) for modeling and testing data (260) for validation. Each data set has the same pattern of good/bad debtors with the full data set, which comprise of 70% good debtors and 30% bad.
2. Data exploration.
3. Modeling the data by using stepwise, forward, and backward logistic regression. The probability modeled was \(Y=1\) (the debtor had a good collectability status). Then choose one of those three models by considering the fit of the model and the model having the highest c statistic.
5. Determine optimal cutpoint from the intersection of sensitivity and specificity.
6. Model validation with testing data.
7. Comparing the classification rate and the c statistic between logistic ridge regression and the logistic regression with variable selection.
8. Generate V2* that had some specific correlation with V1. Then do step 3 until step 7 with new data (by replacing V2 with V2*) to see the performance of logistic regression with variable selection and logistic ridge regression as the correlation between V1 and V2* increases.

**RESULT AND DISCUSSION**

**Data Exploration**

There were no outliers and missing values in the full data set, so all of the 1000 observations were included in the analysis. Allocation of the data into modeling and validation was based on the proportion of bad and good cases of the overall data set. Each had 70% of good and 30% of bad which was appropriate with the full data set.

The variables of V1 (duration of credit) and V2 (credit amount) had a decreasing trend to the response variable. Figure 4 showed that as the amount of the credit increased, the proportion of debtors with good collectability status decreased. The debtors with high installment rate (V4) tend to be bad debtors. The difference of good debtors for each occupation category was not significant. The group of debtors who were unemployed/unskilled-nonresident had the highest proportion of good debtor compared to the unskilled-resident, official, and officer.

[Figure 4: Plot of percentage of good debtors in each group of credit amount (V2)](attachment:figure4.png)

It can be seen in Appendix 2 that based on age (V3), the group of debtor aged 20 years old until 50 years old had a positive trend to the proportion of good debtor. As the age