Loan Status Prediction

Team members: Jinjin He, Canyang Jin, Huimeng Zhang

- **Abstract:**

We aim to decrease loss of interest of investors and keep up the reputation of P2P. We firstly apply logistic regression on the training set to forecast loan status of debtors, focus on true positive rate to avoid high ratio of bad loans in the lending club system. Then we apply multiple linear regression to forecast their total payments if they do not fully pay their loan. During this process, we also try several methods to solve potential problems in the dataset and evaluate these models' performance through the test set.

- **Introduction:**

The peer-to-peer(P2P) lending has become more and more popular recently. It is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. By using the data provided by the Lending Club, a site that brings investors and borrowers together to put together loans that will benefit both parties, we want to make prediction of the repayment amount of the debtors who has high probability of changed off. We attempt to achieve our goal by two steps:

(1) Predict the future loan status (fully paid vs charged off) by building the logistic regression model.

(2) Predict the repayment of the debtors whose predicted loan status is charged off by building the linear regression models.

- **Methods and Materials:**

We found the dataset on Lending Club website (https://www.lendingclub.com/info/download-data.action). In our raw datasets, there are 42,532 instances and 41 attributes. It contains information on credit grade, annual income, home ownership, total payment, purpose, loan status, etc. After removing uninformative attributes, such as ID number and address, our dataset contains 25 attributes. There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. Sometimes a model has a relatively small training MSE, but a large test MSE because of overfitting. So, it is not appropriate to evaluate a model by training MSE. To avoid this situation, we need to split dataset.

Using stratified sampling, we now split our dataset into a training set, containing 37202 observations, and a test set, containing the remaining 5189 observations. We fit a model on the training set and then evaluate this model on the test set.

Our project aims to capture debtors who won't fully pay their loan and their total payments. Firstly, we fit a logistic regression model on the training data to predict whether a debtor will fully pay its loan and evaluate its performance on the test data. Secondly, we fit a linear regression model to predict the total payment they would pay and evaluate its performance by test MSE.

(1) Logistic regression:

In this step, we want to make prediction of loan status of debtors.

a. multicollinearity remedial measures

We detect collinearity by looking at the correlation matrix of the predictors. If an element of this matrix is large in absolute value, it indicates a pair of highly correlated variables and therefore a collinearity problem in the data. However, collinearity may exist between three or more predictors even if no pair of predictors has a particularly high correlation. So, we compute the *variance inflation factor* (VIF). As a rule of thumb, a VIF value that exceeds 10 indicates a problematic amount of collinearity. (James, Wittern, Hastie, & Tibshirani, 2017, p.101) We removed four predictors: funded_amnt, addr_state, installment and int_rate.

b. Feature Selection

Since we still have 20 predictors, the number of possible models 2^p is still a huge number. Evaluating all the possible alternatives is daunting endeavor. We apply forward stepwise selection to compare with the various all-possible-regressions procedures and find out the appropriate subset based on AIC criterion, as shown in the fig 1.

According to the plot of number of predictors and AIC values, we choose four of them (collection_recovery_fee, last_pymnt_amnt , grade, debt_settlement_flag) for our logistic regression model. The description of these 4 predictors is shown in the table 1.

In the training set, there are 31650 observations in class “fully paid” and 5552 observations in class “charged off”. Obviously, the “fully paid” class is significantly bigger than the “charged off” class. In this situation, classification algorithms tend to favor the large class. In our project, the person whose predicted class is “fully paid”

and actual class is “charged off” would hurt the investor’s interest. Since our data has class imbalance problem, *true positive rate* (TP rate) is more suitable to evaluate our model than accuracy rate. Since there exists a tradeoff between the accuracy rate and TP rate, we intend to get a higher TP rate with a moderate accuracy rate. For example, the probability of loan_status = “charged off” given predictors’ values can be written as Pr(loan_status= charged off| predictors). The values of Pr(loan_status= charged off | predictors), which we abbreviate P(predictors), will range between 0 and 1. Then for any given value of predictors, a prediction can be made for loan_status= charged off.(James, Wittern, Hastie, & Tibshirani, 2017, p.131)For example, one might predict loan_status= charged off for any individual for whom P(predictors) > 0.5. Since we wish to be conservative in predicting applicants who are at risk for default, we try a few smaller thresholds, such as P(predictors) > 0.3. The summarized result is shown in the table 2.

When we set threshold to 0.1, the corresponding TP rate and accuracy rate are 86.88% and 81.90%, respectively. Further decreasing the threshold leads to a relative low accuracy rate. Thus, we set 0.1 as threshold.

(2) Linear regression model

After we figure out debtors who won’t fully pay their loan in the first step, we further wonder how much they would pay. The more they pay, the smaller the damage degree of investors. After investors know the probability that a certain debtor will not fully pay its loan, they are further wondering how much he or she would pay. Thus, our second step is to predict the total payment of debtors in class “charged off”.

We fit a polynomial linear regression model to this training data to predict Y. Many problems may occur, and we mainly consider 5 most common potential problems.

a. Collinearity

b. Overfitting

c. Violations of model assumptions.

d. Non-linearity of the response-predictor relationships.

e. Outliers, high-leverage points, and influential cases.

In this step, we select observations with loan status = "charged off" from the training data in the first step as our new training data in the second step. There are 5552 instances and 24 predictors. The response is the total payment of a debtor. We have 5 steps to solve these 5 problems,

a. multicollinearity remedial measures

Since we change our response, we need to redo VIF. In this step, we delete 4 predictors, which are funded_amnt, addr_state, loan_amnt, and grade. Now all the VIF values of attributes in our training set are less than 10.

Furthermore, ridge regression can also solve collinearity, adding penalty if we include many attributes and allowing biased estimators of the regression coefficients. (James, Wittern, Hastie, & Tibshirani, 2017, p.215) It has the effect of shrinking estimates of coefficients toward zero. So, it is effective when dealing with multicollinearity problem. As tuning parameter λ increases, the flexibility of this model decrease. It is important to choose a suitable value for the tuning parameter.

Based on cross-validation fig 2, we select optimal tuning parameter, with the value of 0.37649.

b. Feature Selection

Feature selection is finding a subset of good attributes. Feature extraction may construct new attributes, as with principal components analysis (PCA). We have 20 predictors now. To solve the second problem, overfitting, we apply forward stepwise selection to create a subset of predictors based on adjusted R^2, C_p, AIC, and BIC. The BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller subset of attributes. From the fig 3, adjusted R^2, C_p and BIC choose 19, 17, and 8 attributes, respectively. Since we want a much simpler model to avoid overfitting, we choose 8 predictors based on BIC.

Then we fit least square regression model with 8 predictors to make prediction for Y. The adjusted R^2 is 52.88%, indicating that 52.88% of the total variability in the dataset can be described through this considered model.

$$\begin{aligned}\widehat{Y_i} = -5944 + 0.2698 \cdot funded\ amount + 51.81 \cdot term + 10.46 \\ \cdot installment + 0.005826 \cdot annual\ interest - 983.0 \\ \cdot purpose - 143.3 \cdot inquiries + 2.84 \cdot recovery\ fee \\ + 4591 \cdot debt\ settlement\ flag\end{aligned}$$

The description of these 8 predictors is shown in the table 3.

There is another method of feature selection, PCA. PCA finds eigenvectors of correlation matrix. Based on cross-validation fig 4, from the plot we select 4 principle

components. PCA emphasizes the directions of maximum variability, which may not be related to the structure that we are interested in. The MSE is 23984659, which is not ideal.

c. Non-constant variance and normality of error terms

Model assumptions include that the error terms ε_i are normally distributed and the error terms ε_i have a constant variance. If we need to solve violations of normality assumptions, we can apply several kinds of transformations on Y. If we need to solve violation of unequal error variances, we can apply transformation and weighted least squares. Thus, in this part, we check the normality assumption and constant variance assumption, and we try both transformation on the response and weighted least squares regression if there exists any violation of model assumptions. Firstly, we plot residuals vs fitted values. We can see a funnel shape in the fig 5, indicating that the variance of residuals increases with the value of the fitted values. Then we conduct Breusch-Pagan test and Lilliefors (Kolmogorov-Smirnov) normality test. The p values of these two tests are both less than 2.2×10^{-16}, indicating violations of these two model assumptions.

To deal with that, we try different concave transformations of the response, such as $\log Y$, $\sqrt{Y}$, $Y^{0.1}$. Among all these transformations, the best p-value for Lilliefors (Kolmogorov-Smirnov) normality is 0.03931 with the response $\sqrt{Y}$. Therefore, our response becomes $\sqrt{total\ payment}$ now.

However, all p-values for Breusch-Pagan test are all less than 2.2×10^{-16}. So, we use weighted least squares regression to modify this problem. We try several kinds

of weight functions to estimate the regression coefficients. According to training MSE values, we select $\frac{1}{\hat{y}}$ as weight function, and the corresponding test MSE is 557.0247, which is pretty good. We can also see that this weight function has the smallest test MSE value, indicating that we don't have overfitting problem in our model after we remove redundant attributes by VIF values.

d. Non-linearity of the response-predictor relationship.

Although a linear regression model assumes that the relationships between the response and attributes are linear, in real world dataset this assumption is hard to be satisfied. So, we also include interaction terms. Then we get numerous transformed attributes. In order to avoid overfitting, we apply forward stepwise selection again. From this BIC plot (fig 6), we select 12 attributes.

We consider a multiple linear regression with 12 predictor variables, as given in:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{12} X_{i12}$$

The estimated function is:

$$\begin{aligned}\widehat{\sqrt{Y_i}} = -26.71 &+ 0.09269 \cdot x_1 + 0.01867 \cdot x_2 + 7.633 \times 10^{-5} \cdot x_3 \\ &- 1.905 \times 10^{-6} \cdot x_4 - 4.237 \times 10^{-4} \cdot x_5 - 1.437 \times 10^{-4} \cdot x_6 \\ &- 3.169 \times 10^{-4} \cdot x_7 - 2.127 \times 10^{-3} \cdot x_8 + 5.236 \times 10^{-7} \cdot x_9 \\ &+ 4.227 \times 10^{-4} \cdot x_{10} + 0.07658 \cdot x_{11} - 0.002072 \cdot x_{12}\end{aligned}$$

Then we begin analysis of variance by table 4.

$SSE(X_1)$ measures the variation in Y when X_1 is included in this model.

$SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in this model.

$$SSR(X_1) = 3281752$$

$$SSR(X_2|\, X_1) = 182450$$

$$SSR(X_3|X_1, X_2) = 445478$$

$$R^2_{Y1} = \frac{SSR(X_1)}{SSTO} = \frac{3281752}{4204620 + 3281752} = 43.84\%$$

It accounts for 43.84% of variation of Y explained by X_1.

The relative marginal reduction in the variation in Y associated with X_2 when X_1 is already in the model is:

$$R^2_{Y\,2|1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{182450}{3281752} = 5.56\%$$

It accounts for the coefficient of partial determination between Y and X_2, given that X_1 is in the model.

$$R^2_{Y\,3|21} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{445478}{4022170} = 11.08\%$$

It accounts for the coefficient of partial determination between Y and X_3, given that X_1 and X_2 are in the model.

Next, we test whether X_3 can be dropped from this regression model given that X_1 and X_2 are retained. We set level of significance $\alpha = 0.05$.

The null hypothesis and its alternative:

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

The test statistic:

$$\mathrm{F}^* = \frac{SSR(X_3|X_1, X_2)/1}{SSE(X_1, X_2, X_3)/(n-4)} = \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} = \frac{445478}{645} = 690.66$$

Decision rules: If $\mathrm{F}^* \leq F(0.95; 1{,}5548) = 3.843135$, we conclude H_0. Otherwise, we conclude H_1. Since $\mathrm{F}^* = 690.66 > 3.843135$, we conclude H_1, which indicates

that given that X_1 and X_2 are retained, X_3 cannot be dropped from the regression model.

e. Remedial Measures for influential cases

We measure the influence of these outlying cases on the fitted values and estimated regression coefficients by means of DFFITS, Cook's Distance, DFBETAS, COVRATIO, and beyond. We delete 541 influential cases among 5552 observations. Although the adjusted R^2 is improved, p values for normality and constant variance seems to be much worse. Summarized results are shown in the table 5.

Breusch-Pagan test for the model after we delete influential cases: This test assumes that the error terms are independent and normally distributed and that the variance of the error is related to the level of X in the following way:

$$log_e \delta_i^2 = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + +\gamma_{12} X_{12i}$$

The null hypothesis and its alternative:

$$H_0: \text{all } \gamma_i' s \text{ are equal to zero}$$

$$H_a: \text{o. w.}$$

We set level of significance $\alpha = 0.05$.

Test statistic: $X_{BP}^2 = \frac{SSR^*/2}{(\frac{SSE}{n})^2} = 586.56$

Decision Rule: If $X_{BP}^2 > X_1^2(1-\alpha)$, reject H_0;if $X_{BP}^2 \leq X_1^2(1-\alpha)$, we fail to reject H_0. Since $X_{BP}^2 = 586.56 > 21.02607$, we reject H_0. So, we conclude that the variance of the error term is not constant.

Lilliefors (Kolmogorov-Smirnov) normality test for the model after we delete influential cases:This test assumes distribution parameters known. We estimate μ, δ^2 from sample.

The null hypothesis and its alternative:

$$H_0: \text{sample come form } \mathrm{N}(\mu, \delta^2) distribution, \mu, \delta^2 \ unknown$$

$$H_1: \text{o. w.}$$

We set level of significance $\alpha = 0.05$.

Test statistic: $\mathrm{D} = \sup [\mathrm{F(x)} - \mathrm{G(x)}]$

Decision Rule: If D exceeds the critical value or p-value is less than α, reject H_0; Otherwise, we fail to reject H_0. Since p-value is less than 0.05, we reject H_0. So, we conclude that sample does not come from a normal distribution.

In this case, we also try robust regression to dampen the influence of outlying cases, as compared to ordinary least squares estimation. However, this method does not work, either.

- **Result**

(1) Logistic regression

We set threshold as 0.1 and apply the fitted logistic regression model with four attributes selected by forward stepwise selection on the test set.

In the test set, there are 4422 observations in class “fully paid” and 767 observations in class “charged off”. Among 767 debtors who is “charged off”, we successfully capture 86.84% of them and the test accuracy rate is 82.64%, which is acceptable. Contagious table is shown in the table 6.

(2) Linear regression

In order to solve the violation of constant variance, we use transformations on Y and weighted least squares regression. Among all these transformations, the best p-value for Lilliefors (Kolmogorov-Smirnov) normality test is 0.03931 with the response $\sqrt{Y}$ but Breusch-Pagan test are all less than 2.2×10^{-16} which indicates the violation of constant variance assumption. The corresponding results of weighted least squares regression in shown in the table 7, the test MSE is lower comparing to the previous models but the non-constant variance problem still exists. The corresponding test MSE of the best case with $\frac{1}{\hat{y}}$ as weight function is 557.0247.

We also try robust regression to dampen the influence of outlying cases and summarized results in the table 8, and best p-value of Lilliefors (Kolmogorov-Smirnov) normality test is 0.008273 and all p-values of robust regression models are all less than 2.2×10^{-16}, which indicates the serious non-constant variance problem.

Discussion and conclusions:

We only select 4 predictors in the logistic regression model and the result is pretty good, capturing 86.84% of people who may be charged off. So, collection_recovery_fee, last_pymnt_amnt, grade and debt_settlement_flag are the 4 most important attributes.

With weighted least squares regression, corresponding test MSE of the case with $\frac{1}{\hat{y}}$ as weight function is 557.0247, which is pretty good. the Although we may get a smaller test MSE, all these methods do not work to solve non-constant variance. The

reason may be that the number of observations is too large to remedy the violations of non-constant variance by simply transforming Y globally.

- **Bibliography**

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer, 101, 131, 215.

Appendix: Tables and Plots

Tables:

Table 1: Logistic Regression Attributes Description

	description
collection_recovery_fee	post charge off collection fee
last_pymnt_amnt	Last total payment amount received
grade	LC assigned loan grade
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.
Loan status (response)	Current status of the loan

Table 2: TP Rates with Different Thresholds

Threshold	Train TP Rate	Train Accuracy rate
0.5	68.49%	95.30%
0.3	68.93%	95.08%
0.2	71.63%	93.95%
0.1	86.88%	81.90%
0.08	91.76%	76.15%

Table 3: Linear Regression Attribution Description

	description
Total payment (response)	Payments received to date for total amount funded
Funded amount	The total amount committed to that loan at that point in time.
Term	The number of payments on the loan. Values are in months and can be either 36 or 60.
Installment	The monthly payment owed by the borrower if the loan originates.
Annual interest	The self-reported annual income provided by the borrower during registration.
Purpose	A category provided by the borrower for the loan request.
Inquiries	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
Recovery fee	post charge off collection fee
Debt settlement flag	The most recent date that the Debt_Settlement_Flag has been set

Table 4: Analysis of Variance Table

	predictors	Df	SSE	MSE	F value	P value
$installment$	x_1	1	3281752	3281752	5497.6001	$< 2.2 \times 10^{-16}$
$recovery\ fee$	x_2	1	182450	182450	305.6409	$< 2.2 \times 10^{-16}$
$funded\ amount \cdot term$	x_3	1	445478	445478	746.2654	$< 2.2 \times 10^{-16}$
$installment \cdot funded\ amoun$	x_4	1	73852	73852	123.7174	$< 2.2 \times 10^{-16}$
$funded\ amount \cdot purpose$	x_5	1	6984	6984	11.6387	0.0006506
$funded\ amount \cdot inquiries$	x_6	1	33505	33505	56.1279	7.860×10^{-14}
$funded\ amount \cdot debt\ settlemen$	x_7	1	99814	99814	167.2090	$< 2.2 \times 10^{-16}$
$installment \cdot term$	x_8	1	9378	9378	15.7095	7.478×10^{-5}
$term \cdot annual\ income$	x_9	1	9246	9246	15.4887	8.402×10^{-5}
$recovery\ fee \cdot term$	x_{10}	1	4167	4167	6.9801	0.0082653
$installment \cdot debt\ settlemen$	x_{11}	1	9517	9517	15.9436	6.610×10^{-5}
$recovery\ fee \cdot debt\ settlemen$	x_{12}	1	23800	23800	39.8697	2.926×10^{-10}
$residuals$	5539	5539	3306466	597		

Table 5: Influential Cases

	Adjusted R^2	lillie.test p-val	bptest p-val
Do not delete influential cases.	0.5338	0.03931	< 2.2e-16
Delete influential cases.	0.5713	0.0002574	< 2.2e-16

Table 6: Contagious Table of Test Set

		true	
		Fully paid	Charged off
predict	Fully paid	3596	105
	Charged off	796	693

Table 7: Weighted Least Squares

Weight	Training MSE	Test MSE	lillie.test p-val	bptest p-val
$\frac{1}{\hat{y}}$	596.512	557.0247	6.14e-05	< 2.2e-16
$\frac{1}{res^2}$	628.7692	607.5463	0.03527	< 2.2e-16
$f(res)$	640.0233	620.44	7.692e-07	< 2.2e-16
$\frac{1}{\lvert res\rvert}$	628.8545	607.5413	0.02514	< 2.2e-16
$\frac{1}{\lvert res_1\rvert}$	635.0558	610.283	7.692e-07	< 2.2e-16
$\frac{1}{\lvert res_2\rvert}$	635.0303	610.2441	5.782e-05	< 2.2e-16
notes	$f(res) = 1/fitted(lm(\lvert res\rvert \sim \widehat{\sqrt{y}}))\text{^}2$ $res_1 = externally\ studentized\ residuals$ $res_2 = weighted\ raw\ residuals$			

Table 8: Robust Regression

	Train MSE	Test MSE	lillie.test p-val	bptest p-val
robust regression psi=psi.bisquare	630.8849	608.7783	0.0003873	< 2.2e-16
robust regression psi=psi.hampel	628.3365	607.4484	0.008273	< 2.2e-16
robust regression psi=psi.huber	629.5456	607.9952	0.001817	< 2.2e-16

Plots:

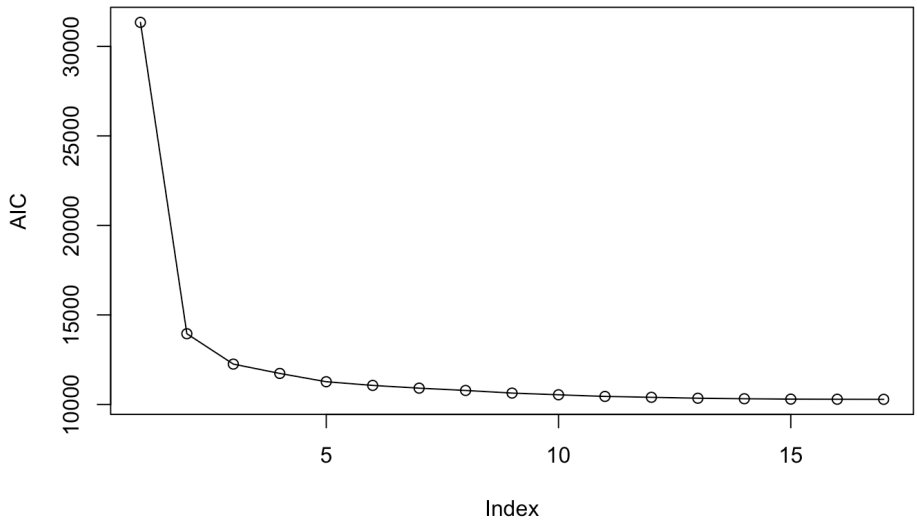

Fig 1: AIC Logistic Regression Plot

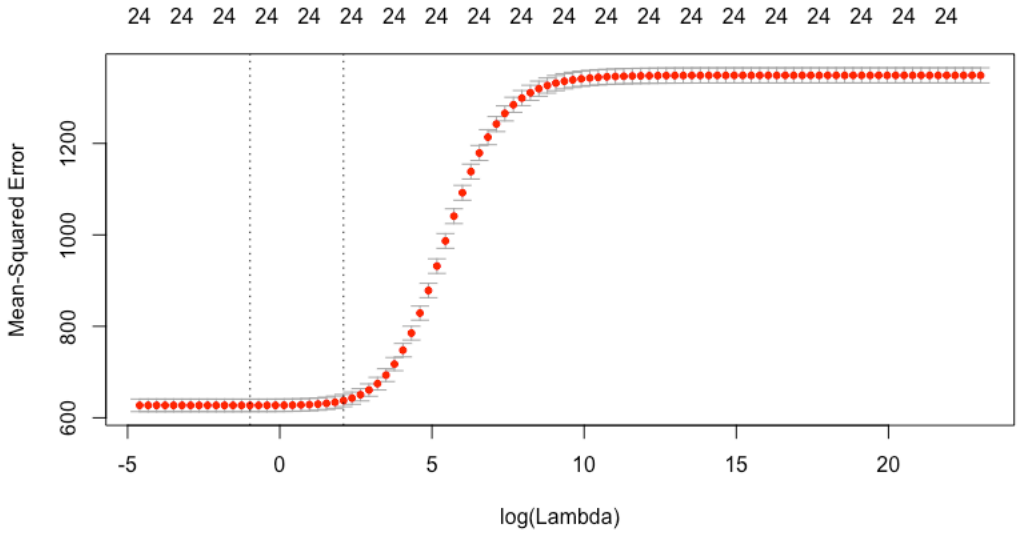

Fig 2: CV Plot of Ridge Regression

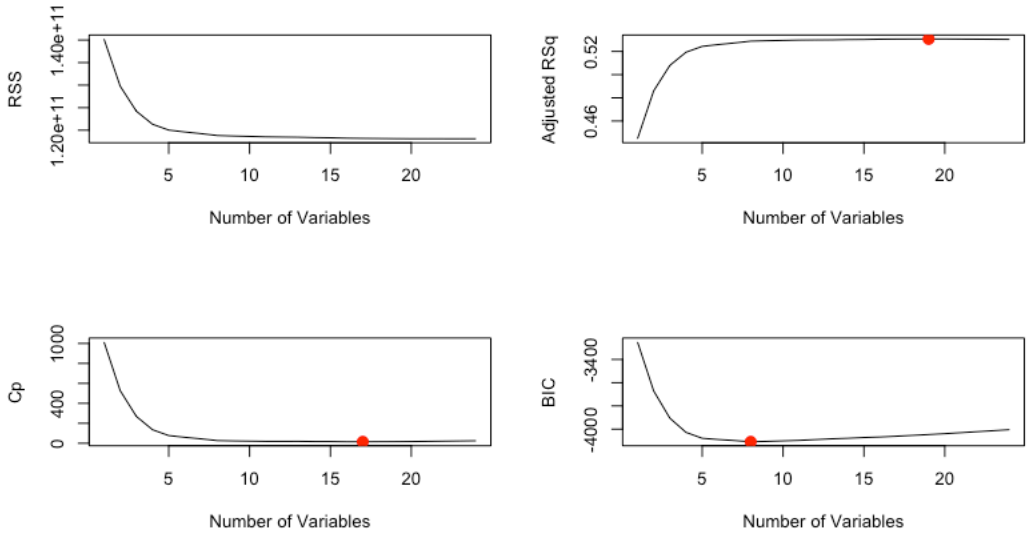

Fig 3: Feature Selection Plots

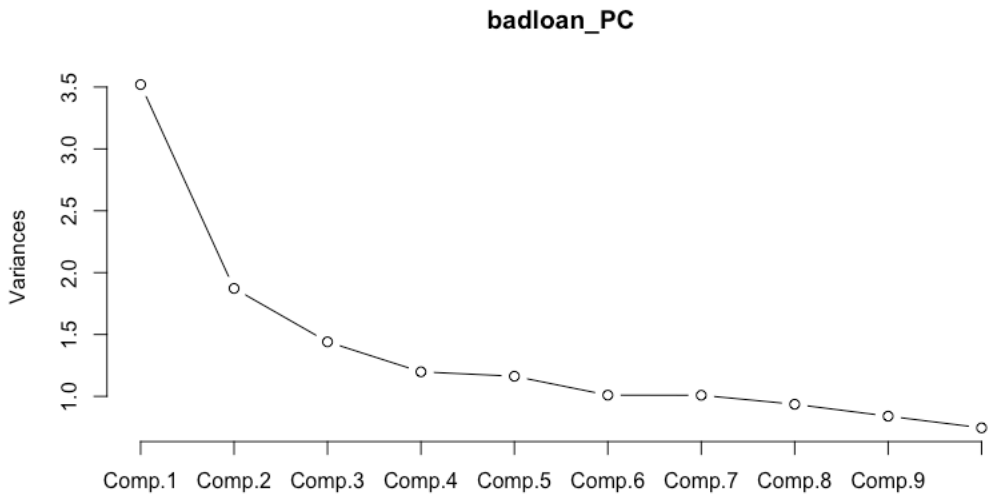

Fig 4: Principle Components

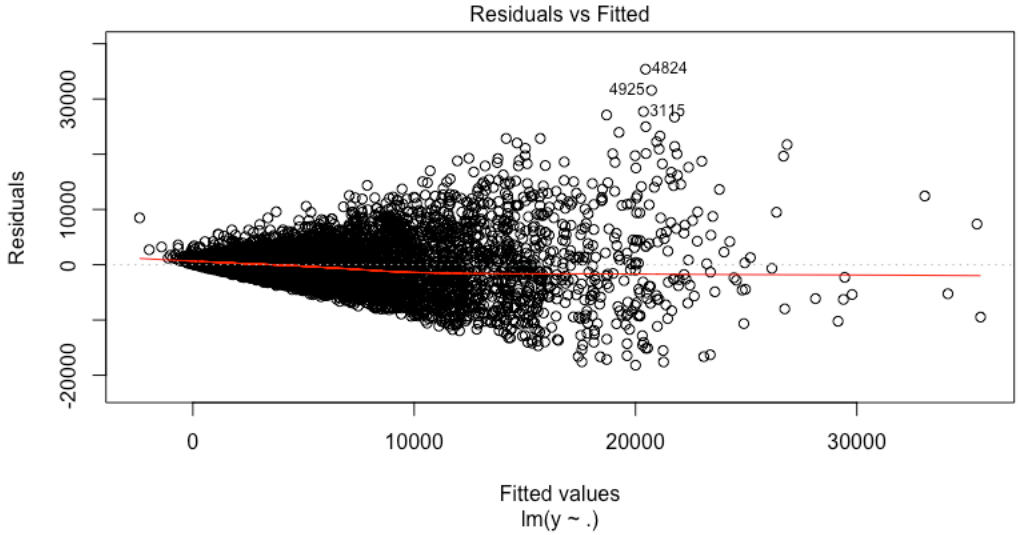

Fig 5: Residuals vs Fitted Values

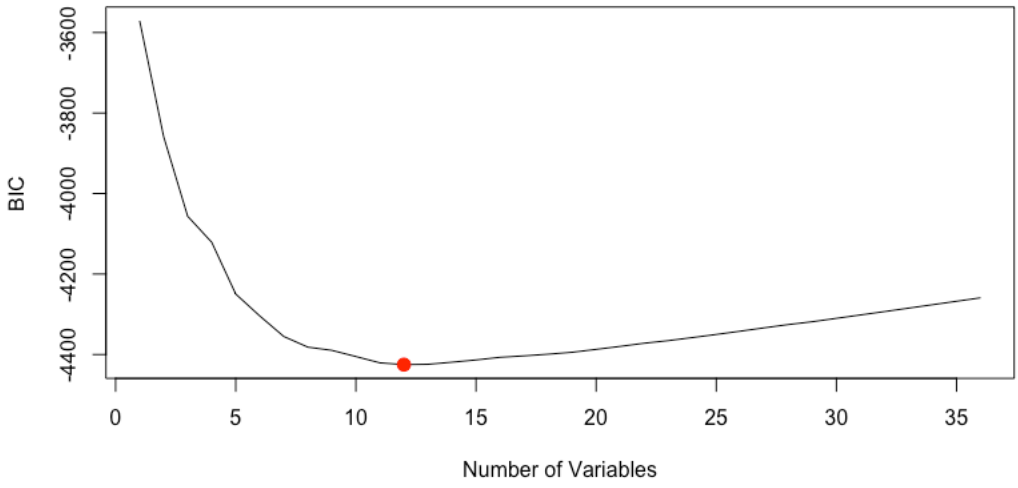

Fig 6: BIC Plot