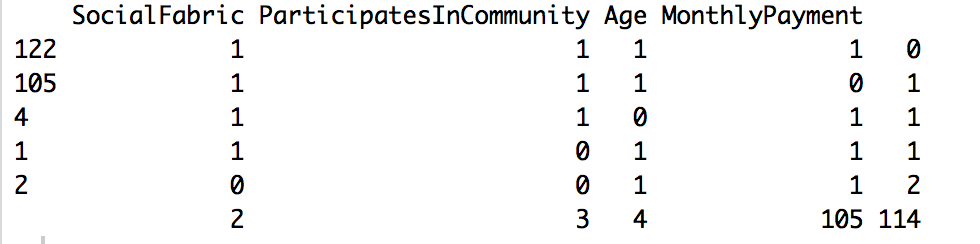
A company wants us to help them use analytics to create a model predicting whether its customers will pay back their loan as a function of information it has about them. The data is in “Homework2.csv”. The independent variables are in column B through O. The goal is to predict Column P, which is 1 if the customers did NOT pay back their loan and 0 if they did.

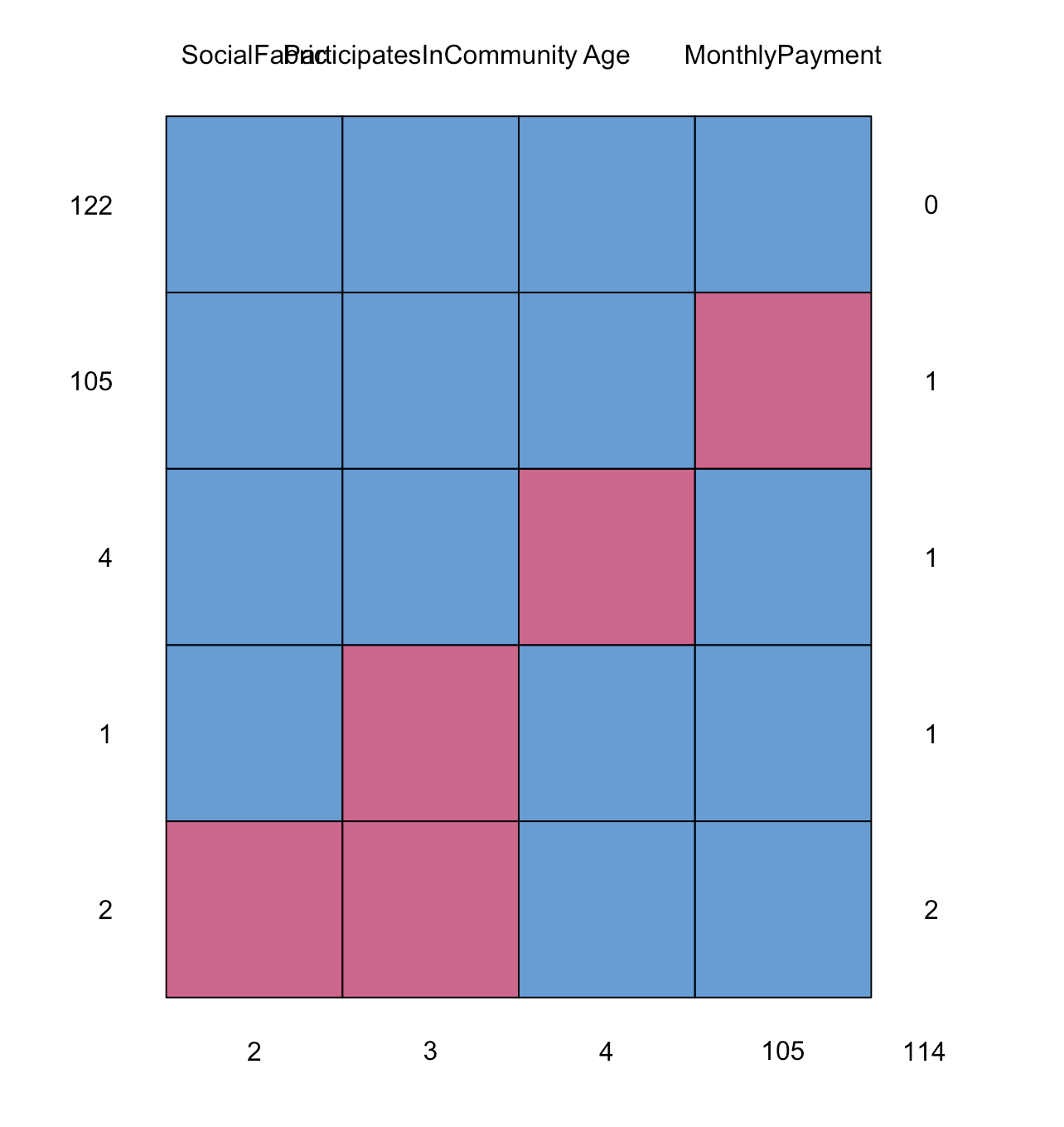
1. Load the mice, VIM and lattice packages. Generate the summary of your data frame using the summary function. Which independent variables have missing values?

* *summary(homework2); missing data:*
  + *Age: 4, SocialFabric:2, MonthlyPayment:105, ParticipatesInCommunity:3*

Create a smaller data frame using only those independent variables and visualize the missing data using the md.pattern and aggr function (this is exactly like mice\_example.R, and the code in LogReg\_MissingData.R shows you how to select a set of columns). What percentage of the data is missing for each feature?

* *simple = homework2[c("Age", "SocialFabric", "MonthlyPayment, ParticipatesInCommunity ")]*
* *summary(simple)*
* *set.seed(88)*
* *md.pattern(simple)*

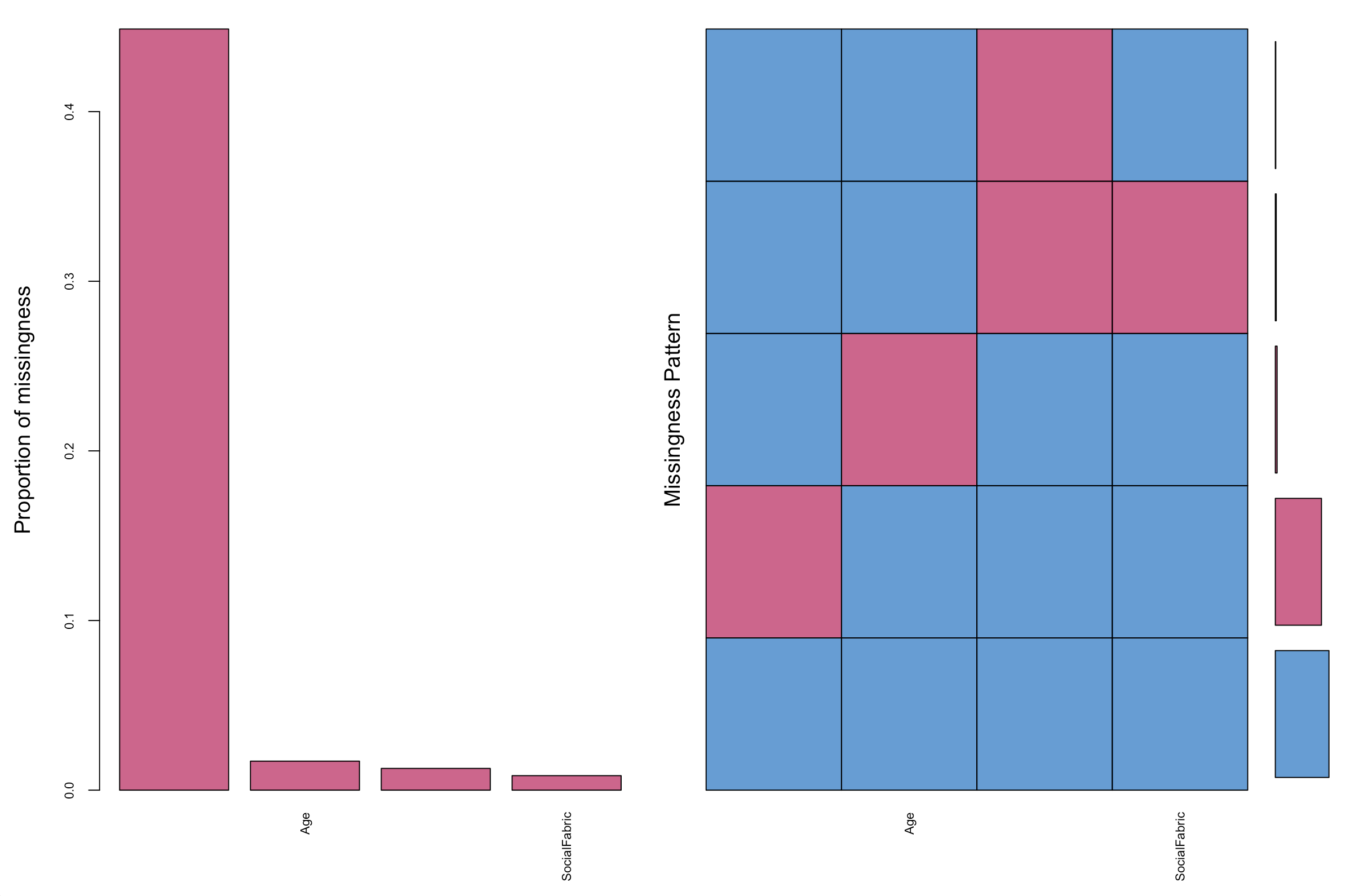




* *simple\_aggr = aggr(simple, col=mdc(1:2), numbers=TRUE, sortVars=TRUE, labels=names(simple), cex.axis=.7, gap=3, ylab=c("Proportion of missingness","Missingness Pattern"))*

*Variables sorted by number of missings:*

* *MonthlyPayment: 0.448717949 ~ 44.8%*
* *Age: 0.017094017 ~ 1.70%*
* *ParticipatesInCommunity: 0.012820513 ~ 1.2%*
* *SocialFabric: 0.008547009 ~ 0.85%*

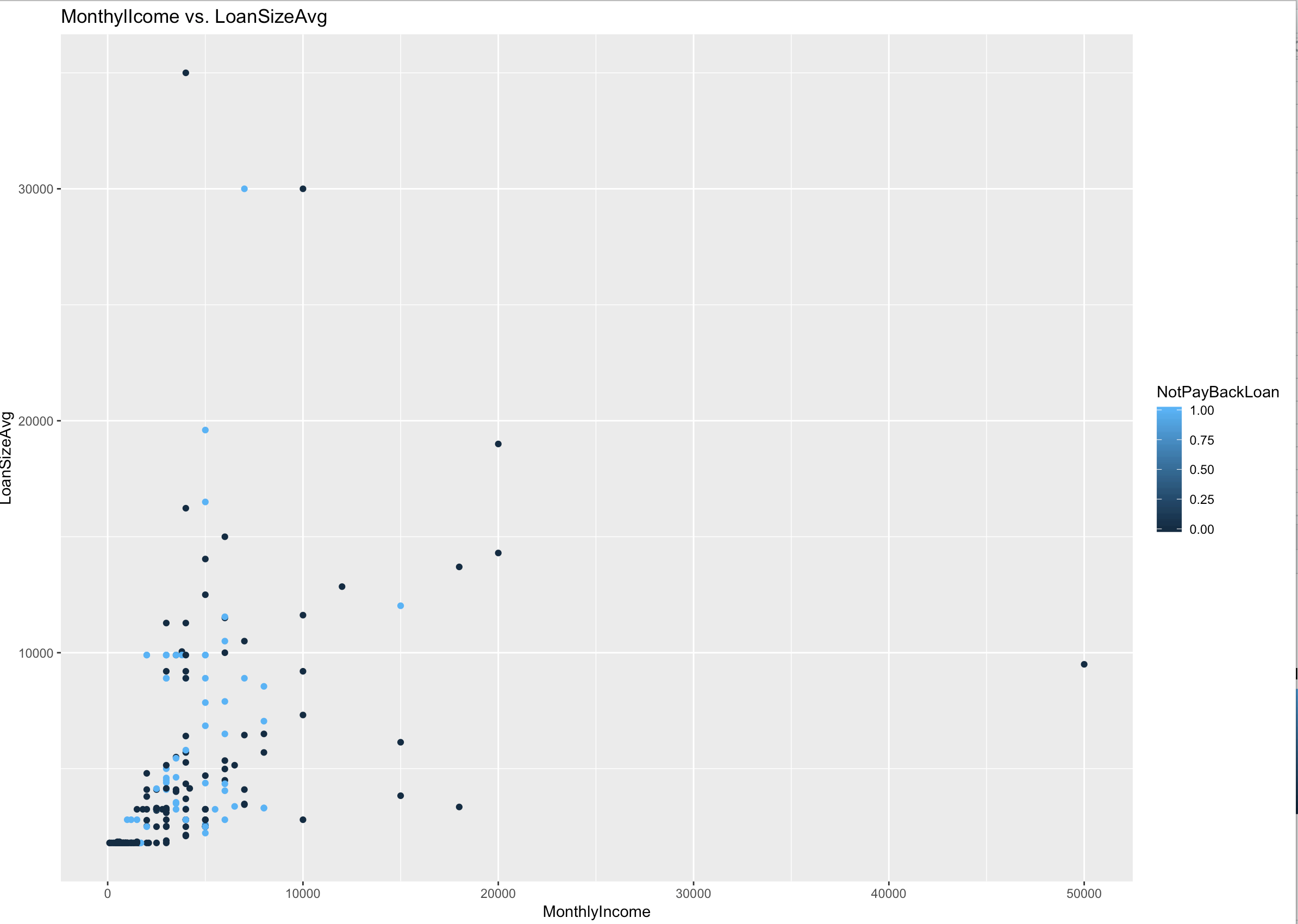


Which independent variable has too many missing data points to be useful and should be deleted? Delete it from your data set.

*MonthlyPayment @ 44.8%*

1. Using ggplot2, create a scatterplot showing LoanSizeAvg as a function of MonthlyIncome, color-coded depending on whether customers paid back their loan or not. Comment on the outliers.

* *scatterplot = ggplot(homework2, aes(x = MonthlyIncome, y = LoanSizeAvg))*
* *scatterplot + geom\_point()*
* *scatterplot + geom\_point(color = "blue", size = 3) + ggtitle("MonthylIcome vs. LoanSizeAvg")*
* *ggplot(homework2, aes(x = MonthlyIncome, y = LoanSizeAvg, color = NotPayBackLoan)) + ggtitle("MonthylIcome vs. LoanSizeAvg") + geom\_point()*

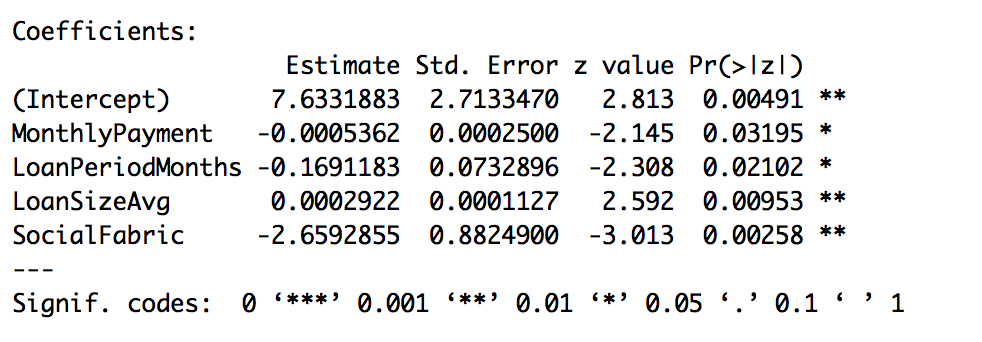


* *MonthlyIncome > 10,000 && LoanSizeAvg > 10,000 are not correlated*

First we will run the logistic regression after removing the instances where there is missing data.

1. Remove the instances where there is missing data using na.omit. How many instances are you left with, and how many did you have in the initial data set? Split the new data set between a training set (65% of the data) and a testing set (35% of the data) and run a logistic regression to predict Column P as a function of Column B through O when missing data is removed. Refine your logistic regression output so that all independent variables are significant at the 90% level or higher, if possible.

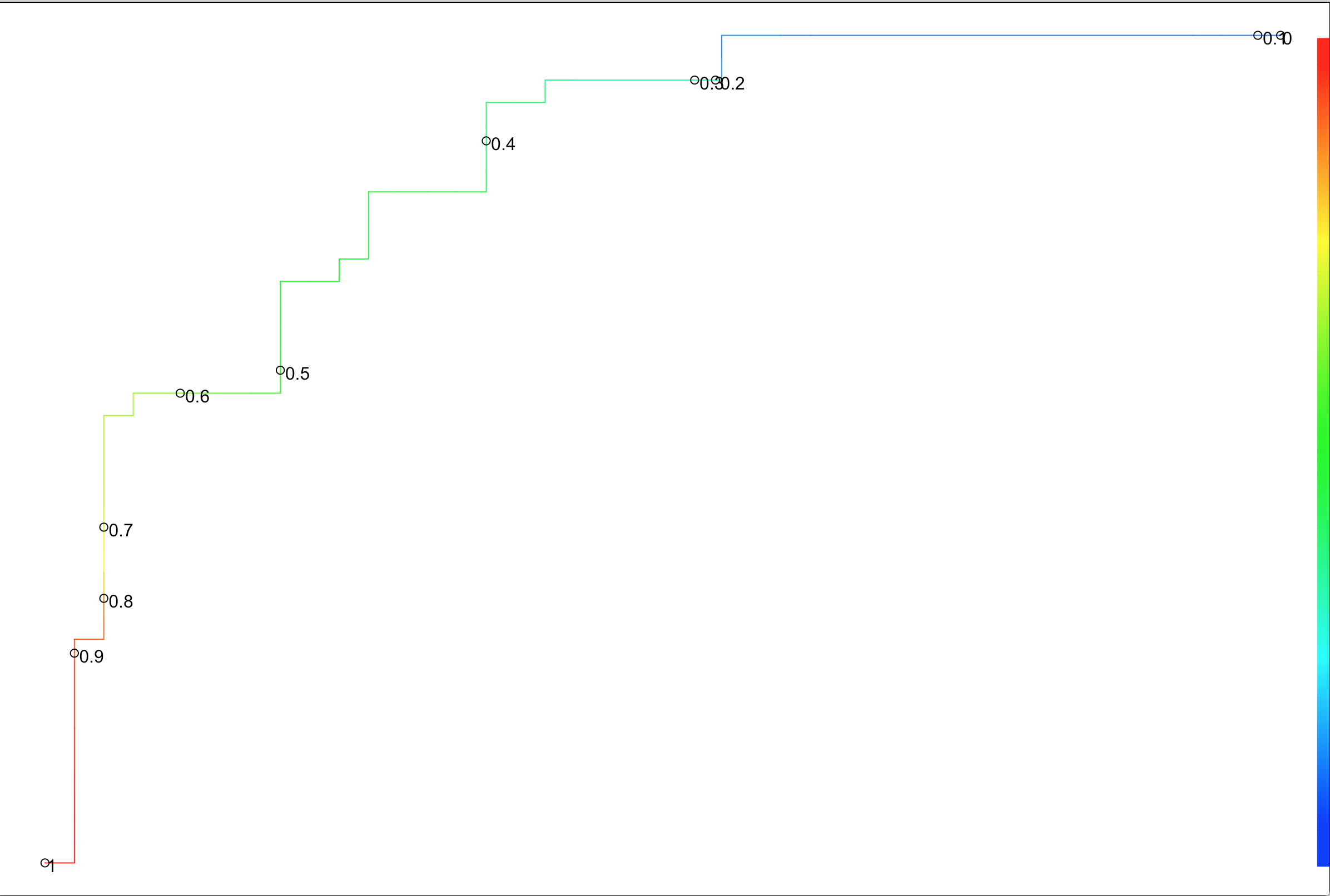
* *homework2\_omit = na.omit(homework2)*
* *summary(homework2\_omit)*
  + *122 observations left with, out of 234 observation we had initially*
* *homework2\_omit\_Train\_Model = glm(NotPayBackLoan ~ MonthlyPayment + LoanPeriodMonths + LoanSizeAvg + SocialFabric, data = homework2\_omit\_Train, family=binomial)*
* *summary(homework2\_omit\_Train\_Model)*



*logit = 7.6331883 - 0.0005362(MonthlyPayment) - 0.1691183(LoanPeriodMonths) + 0.0002922(LoanSizeAvg) - 2.6592855(SocialFabric)*

1. Plot ROC and compute AUC on training set. Which threshold “t” do you recommend to make your predictions? For that threshold, compute the accuracy of your method on your training set.

* *# Accuracy for cutoff = 0.4: (32+27)/(32+27+5+15)= 74.6%*
* *#Sensitivity = 32 /(32+5) = 86.4%;*
* *#Specificity =27/(27+15) = 64.2%; 1 - 0.642 = 35.8% = FPR*
* *# I pick up T=0.4; because Sensitivity > 50% && Specificity > 50%*
* *AUC: 84.6%*



1. Repeat (d) on the testing set (ROC, AUC, accuracy with the same “t” you selected in (d)). Are you satisfied with your model? Why or why not?

*table(homework2\_omit\_Test$NotPayBackLoan, homework2\_omit\_Test\_predict > 0.4)*

*# Accuracy for cutoff =0.4: (18+16)/(18+16+5+4) = 79%*

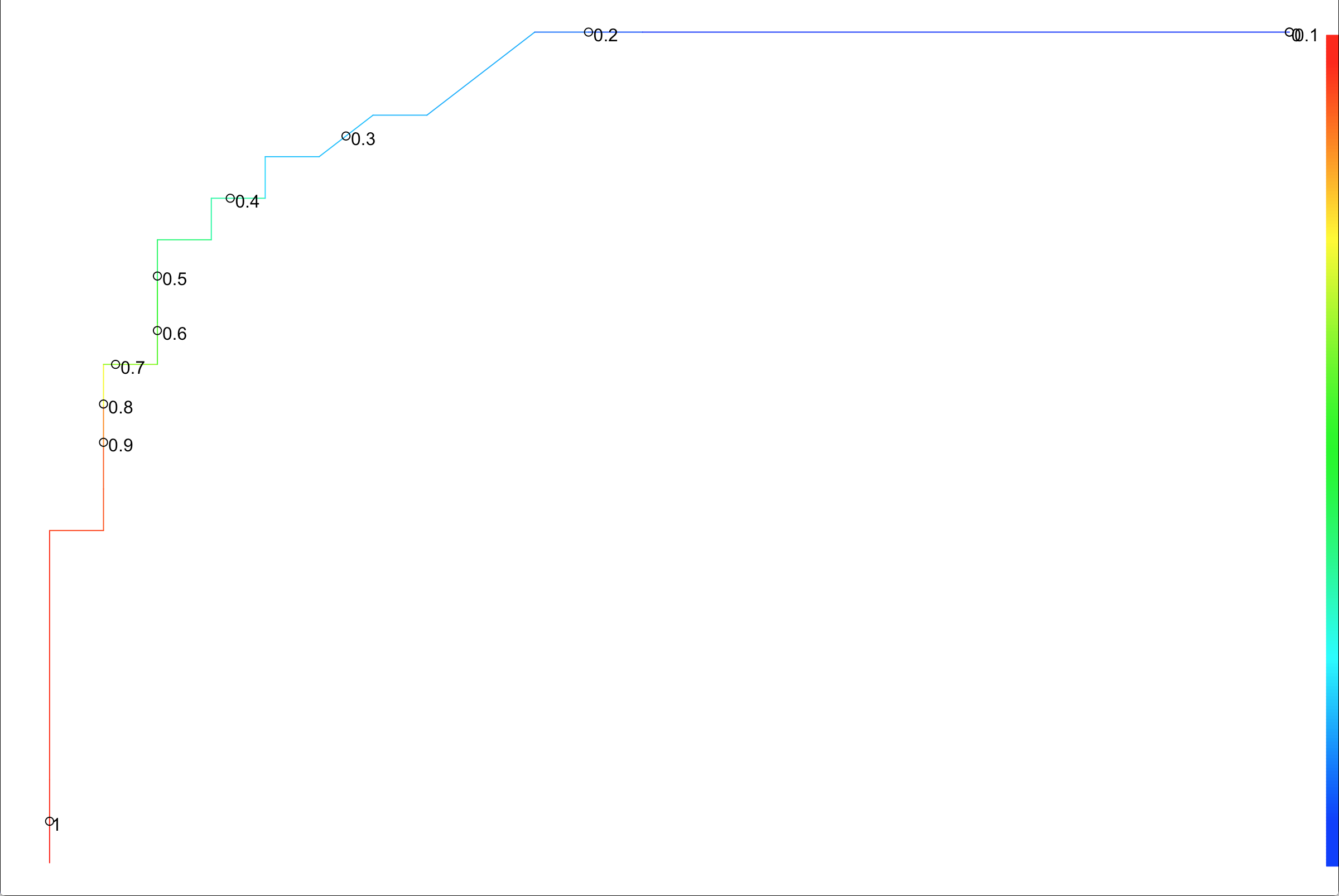
*#Sensitivity = 16/(16+4) = 80%*

*#Specificity = 18/(18+5) = 78.26%; 1 - 0.7826087 = 21.73% = FPR*

*#Baseline = (18+5)/(18+5+4+16) = 53.48%*

*#AUC: 0.9163043 ~ 90.21%. I am not 100% satisfied with the model since it is slightly different from Training set model.*

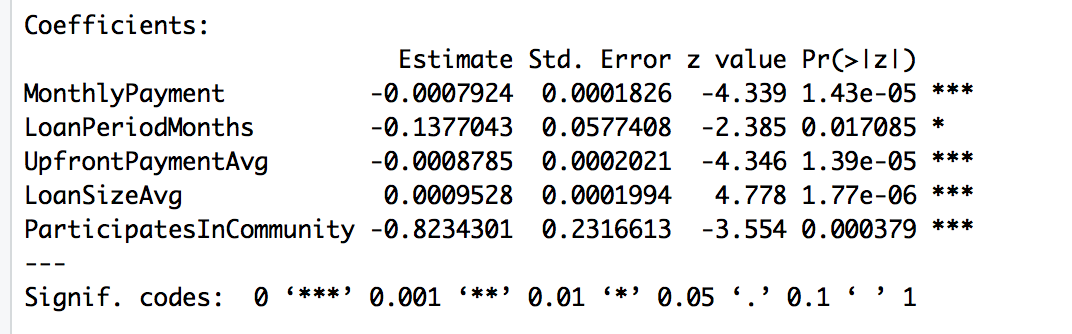
*The AUC, and FPR are slightly different between the Training and the Testing set.*



**(What follows is for EMIS 7357 only – the students in EMIS 5357 will get extra credit if they complete it.)** Then we will run the logistic regression after filling missing data using the “mice” package.

1. Fill in the missing data using the “mice” package (use the parameters m = 5, maxit = 30), using the code seen in class (once you have run the mice function, you need to use the complete function and then fill the missing data in your data set using those values, see LogReg\_MissingData). Split the data between a training set (65% of the data) and a testing set (35% of the data) and run a logistic regression to predict Column P as a function of Column B through O when missing data is filled in using the “mice” package. Refine your logistic regression output so that all independent variables are significant at the 90% level or higher, if possible. Compare the equation you got for the probability of not paying back the loan with the one you obtained when you just discarded instances with missing data.

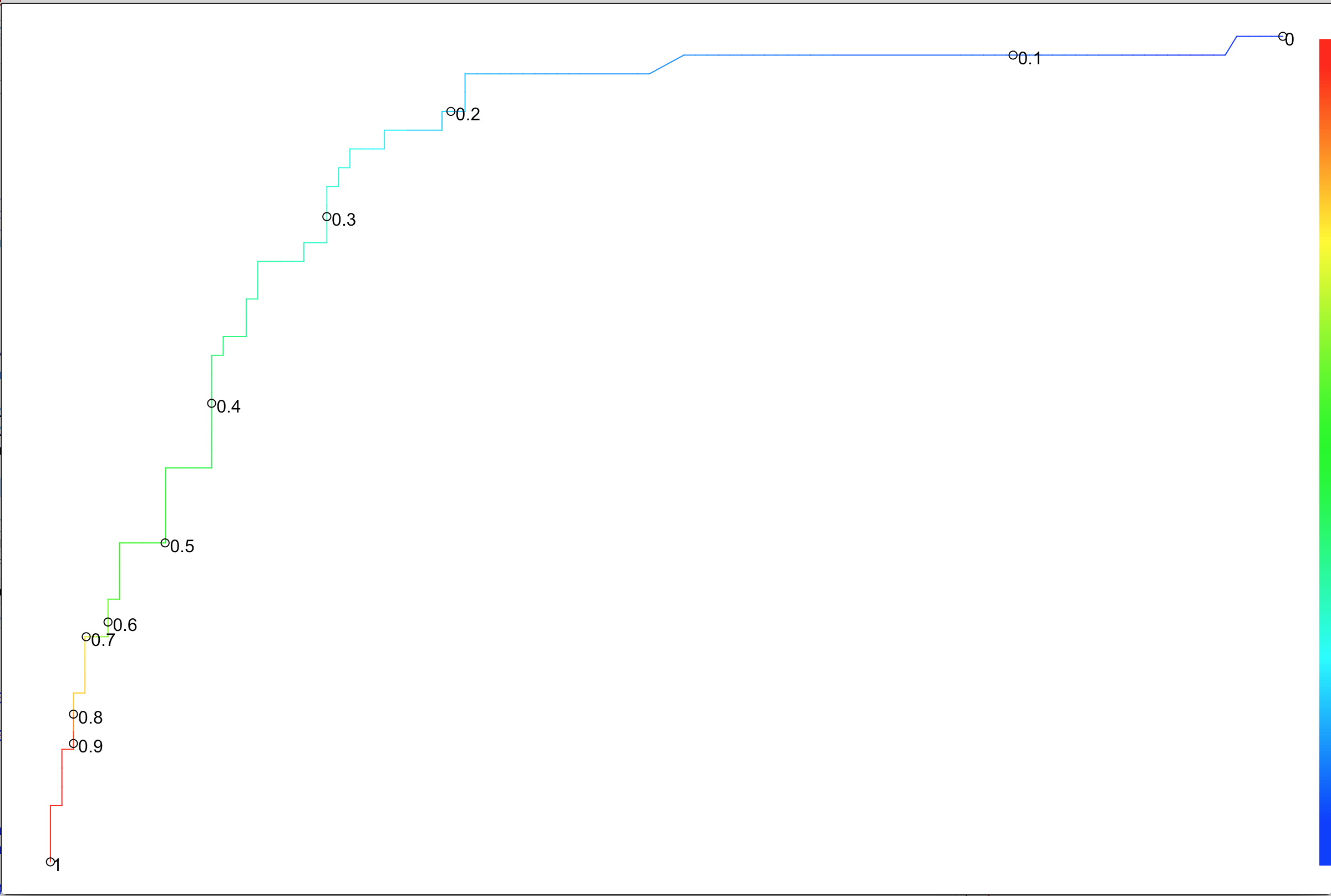
* *#Logit = -0.0007924(MonthlyPayment) - 0.1377043(LoanPeriodMonths) - 0.0008785(UpfrontPaymentAvg) + 0.0009528(LoanSizeAvg) - 0.8234301(ParticipatesInCommunity)*



*The equations are different by some coefficients. For instance, using the mice package does not have a y-intercept, nor socialFabric as coefficient, but it contains UpfrontPaymentAvg.*

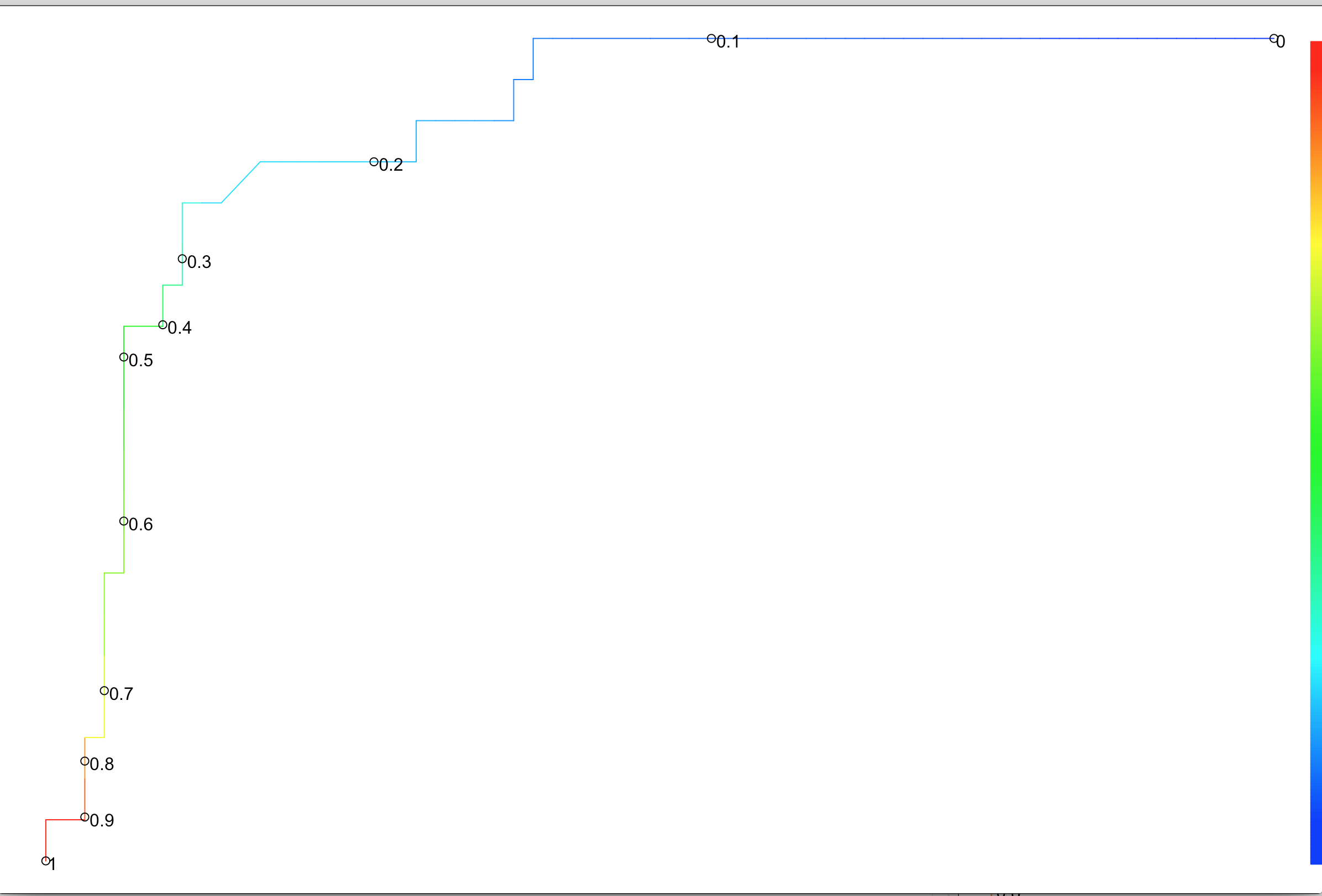
1. Plot ROC and compute AUC on training set. Which threshold “t” do you recommend to make your predictions? For that threshold, compute the accuracy of your method on your training set.

* *table(Train$NotPayBackLoan, pred\_0 > 0.2)*
* *# Accuracy for cutoff = (73+40)/(73+40+34+4)= 0.7483444 ~ 74.8%*
* *#Sensitivity = 40/(40+4) = 0.9090909 ~ 90.9%*
* *#Specificity = 73 /(73+34) = 0.682243; 1 - 0.682243 = 31.7% = FPR*
* *#Baseline = (73+34) /(73+34+4+40) = 0.7086093 ~ 70.8%*
* *# AUC: 0.8532285 ~ 85.3%*

**

1. Repeat (g) on the testing set (ROC, AUC, accuracy with the same “t” you selected in (g)). Are you satisfied with your model? Why or why not? Would you advocate using the model with the missing data omitted or the one with the missing data filled using mice?

* *table(Test$NotPayBackLoan, pred\_0\_test > 0.2)*
* *# Accuracy for cutoff = (47+17)/(47+17+16+3) = 0.77 ~ 77%*
* *#Sensitivity = 17/(17+3) = 0.85 ~ 85%*
* *#Specificity = 47 /(47+16) = 0.7460317; 1 - 0.7460317 = 0.2539683 ~ 25.4% = FPR*
* *#Baseline = (47+16)/(47+16+3+17) = 0.759 ~ 75.9%*
* *# AUC: 0.8904762 ~ 89%*
* *I am satisfied with model since it is similar to the Training set model. Hence, I advocate using this model, with data filled using mice, since the models for Training and Testing set are very similar. The FPR are similar between the Training and the Testing set when using the mice method.*

**

***Grading Scheme***:

EMIS 5357 (a) 1.5 pt (b) 1pt (c) 2.5 pts (d) 1.5 pt (e) 1.5 pt

EMIS 7357 (a) 1pt (b) 0.5pt (c) 1.5pt (d) 0.75pt (e) 0.75pt (f) 2pt (g) 0.75pt (h) 0.75pt