## الف:

درصد وام های بازپرداخت شده:

83.99457

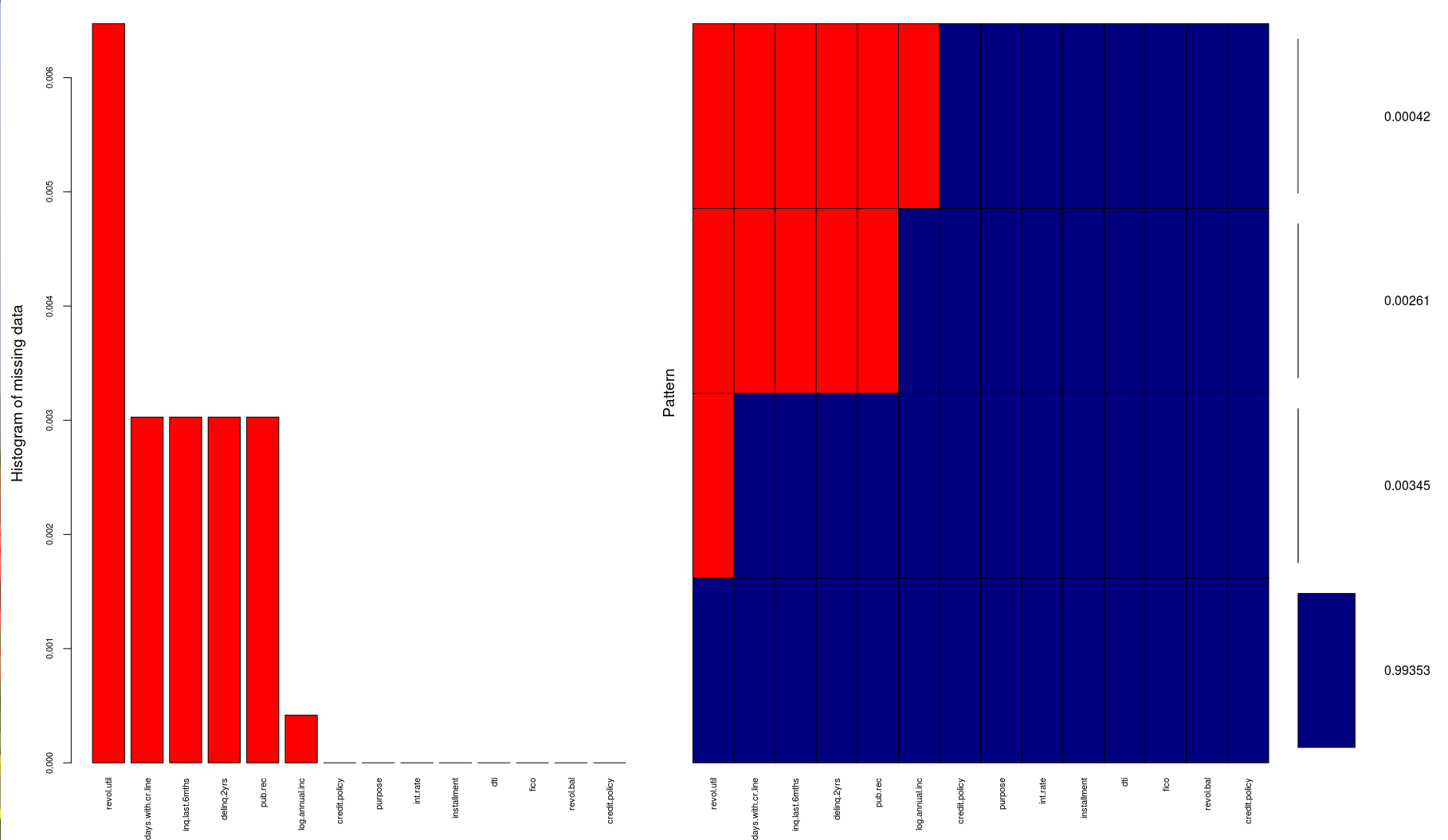
خصیصه های دارای مقادیر مفقود:

log.annual.inc, days.with.cr.line, revol.util, inq.last.6mths, delinq.2yrs and pub.rec

برخورد با داده های مفقود (missing values) در مسائل و زمینه های مختلف، متفاوت است. مثلا اگر مشاهده های دارای مقادیر مفقود، درصد کمی از کل مشاهده ها را شامل شود و به صورت تصادفی در کل داده ها پخش باشد، یعنی مربوط به دسته ی خاصی از داده ها نباشد، می توان از آن ها صرف نظر کرد. اما اگر همین داده های با خصیصه های مفقود شامل اطلاعات مهمی باشند و نتوان از آن ها چشم پوشی کرد روش های مختلفی برای کامل کردن آن ها وجود دارد.  
مثلا می توان به صورت تصادفی مقادیر خصیصه‌ی مفقود را کامل کرد، از میانگین هر خصیصه استفاده کرد، از رگرسیون استفاده کرد (و براساس سایر خصیصه‌ها مقدار را حدس زد)، داده‌ها را بر اساس خصیصه‌ی دارای مقدار مفقود مرتب کرده و مقدار مفقود را با اطلاعات داده‌ی بلافاصله قبل، کامل کرد،و در صورتی که داده‌ها دارای چند خصیصه با مقدار مفقود باشند می‌توان از روش MICE[[1]](#footnote-0) استفاده کرد.

در مورد این سوال، به این دلیل که تعداد داده‌های دارای خصیصه‌های مفقود کم است، می‌توان از آنها چشم پوشی کرد، بدون اینکه باعث ایجاد Overfitting شود. اما از آنجایی که ما می خواهیم بازپرداخت وام را برای داده‌های مفقود و نه فقط برای داده‌های کامل، حدس بزنیم، پس مقادیر مفقود را با استفاده از پکیج mice کامل می کنیم.

باید توجه شود که داده‌های دارای خصیصه های مفقود، نباید بیش از 5 درصد کل داده‌ها را شامل شود. همینطور تعداد خصیصه‌های مفقود هر مشاهده نیز (معمولا)، نباید بیش از 50 درصد از کل خصیصه‌ها، باشد.



شکل فوق درصد داده‌های دارای خصیصه‌های مفقود را نشان می دهد که کمتر از ۱ درصد است.

0.00345 از داده ها، دارای یک خصیصه‌ی مفقود، 0.00261 دارای 5 خصیصه‌ی مفقود و 0.00042 از داده‌ها دارای 6 خصیصه ی مفقود هستند.

برای مشاهده‌ی متدهای قابل استفاده برای کامل کردن داده‌ها، می توان از دستور methods[[2]](#footnote-1) استفاده کرد.

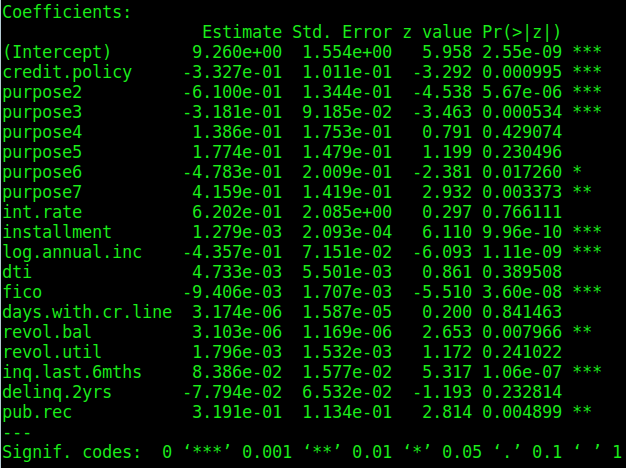
در صورتی که متدی در استفاده از mice انتخاب نشود، با توجه به نوع خصیصه و مقادیر سایر داده‌ها متدی مناسب هر خصیصه، به صورت اتوماتیک انتخاب می شود.

گزینه‌ی مناسب دیگر، می تواند استفاده از Predictive Mean Matching باشد.

## ب:

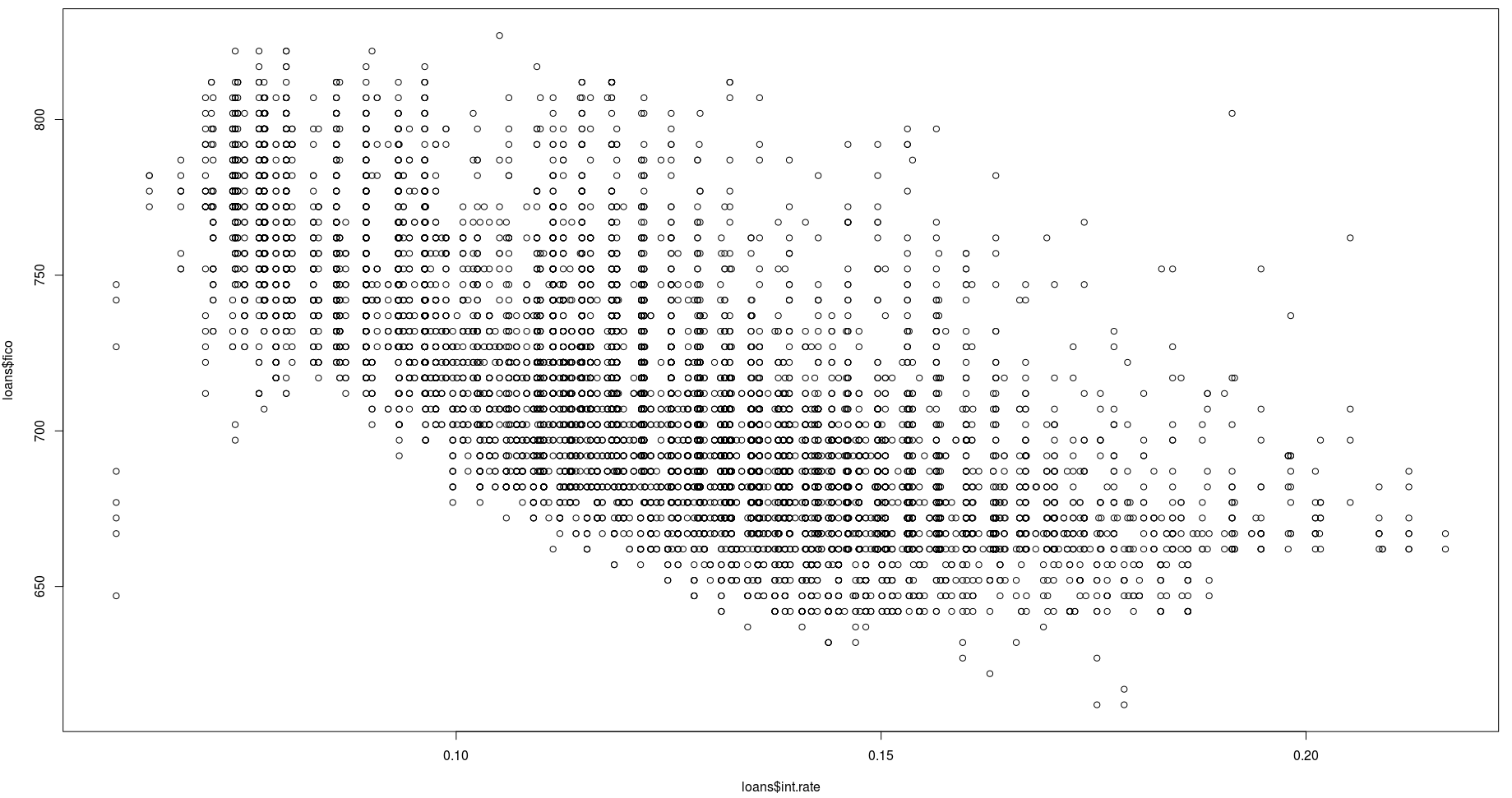
خصیصه های مهم:

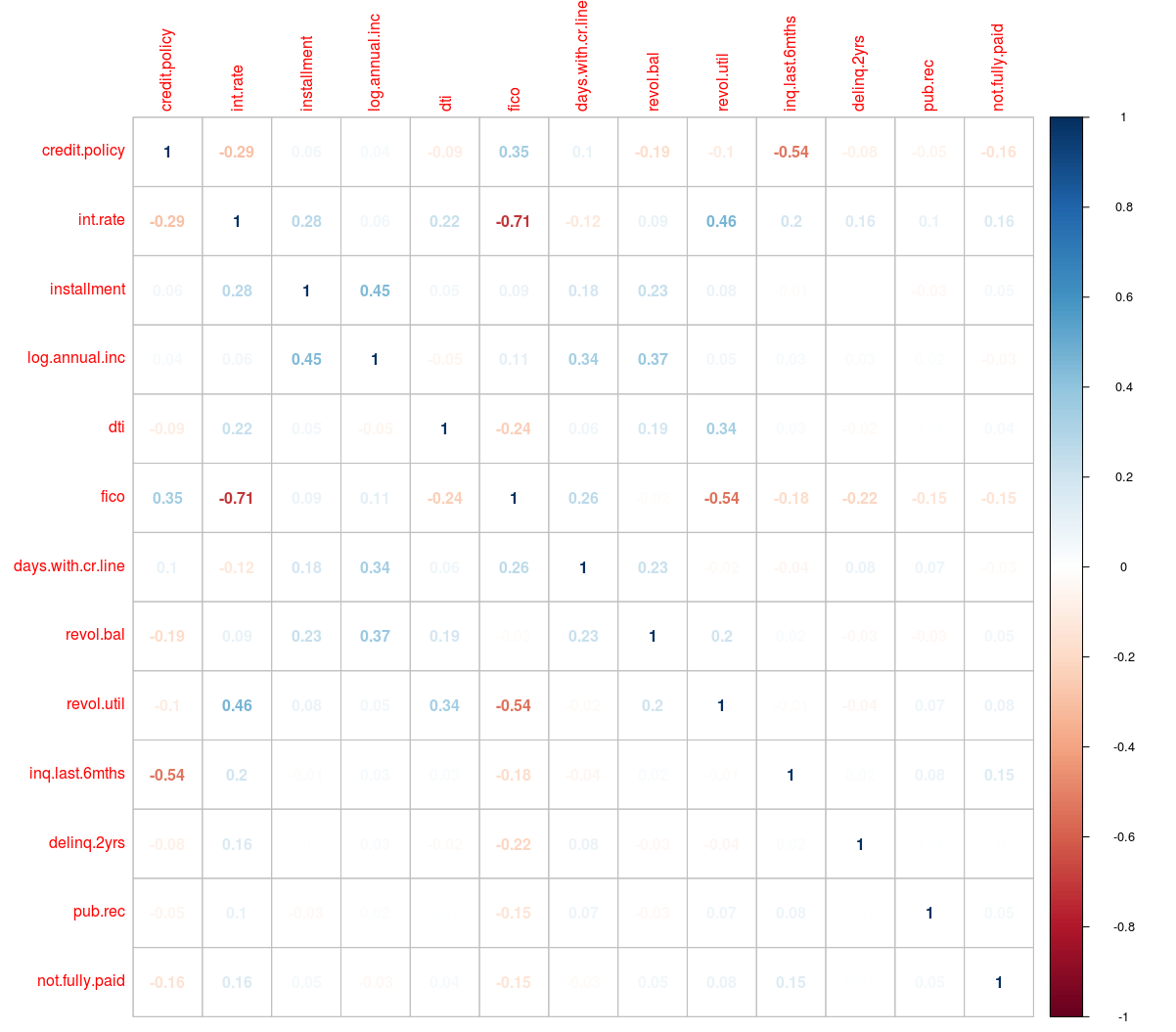
credit.policy ,purpose2,purpose3, purpose6, purpose7 ,installment ,log.annual.inc ,fico,revol.bal,inq.last.6mths ,pub.rec



## ج:

با استفاده از تابع cor می‌توان فهمید که (قدر مطلق) همبستگی int.rate و fico بیش از 0.7 است.





همبستگی خصیصه ها در شکل بالا نشان داده شده است.

با کنار گذاشتن int.rate، دقت مدل از 0.8364 به 0.8368 بهبود پیدا می کند.

همچنین متغییر‌های dti, days.with.cr.line, revol.util, delinq.2yrs نیز تاثیر مثبتی در مدل ندارند و می‌توان آنها را کنار گذاشت.

## د:

**ماتریس سردرگمی**[[3]](#footnote-2) به ماتریسی گفته می‌شود که در آن عملکرد الگوریتم‌های مربوطه را نشان می‌دهند. معمولاً چنین نمایشی برای الگوریتم‌های یادگیری با ناظر استفاده می‌شود، اگرچه در یادگیری بدون ناظر نیز کاربرد دارد.

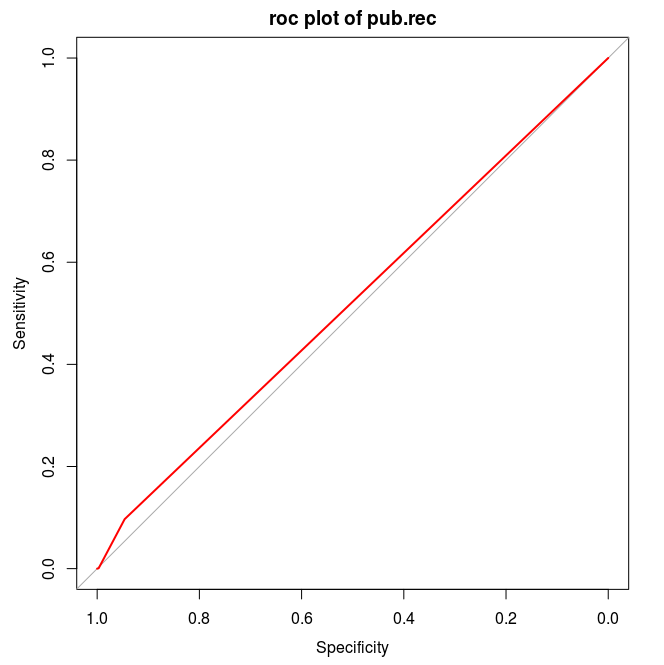
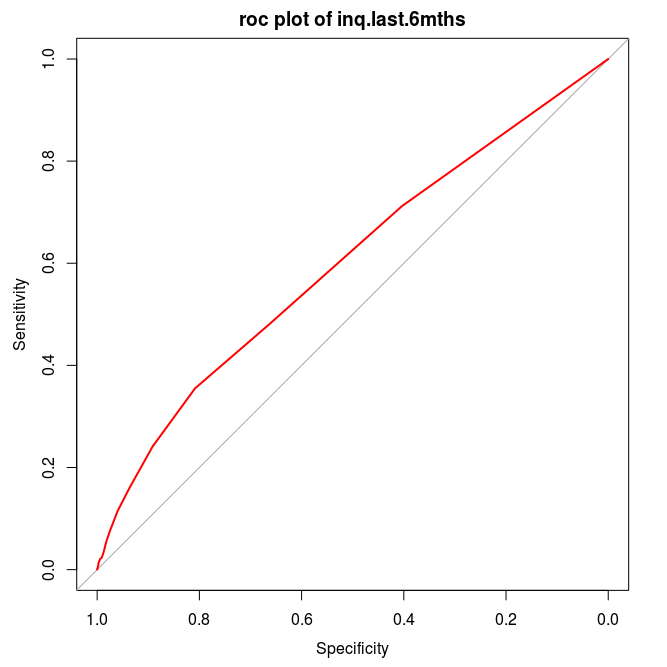
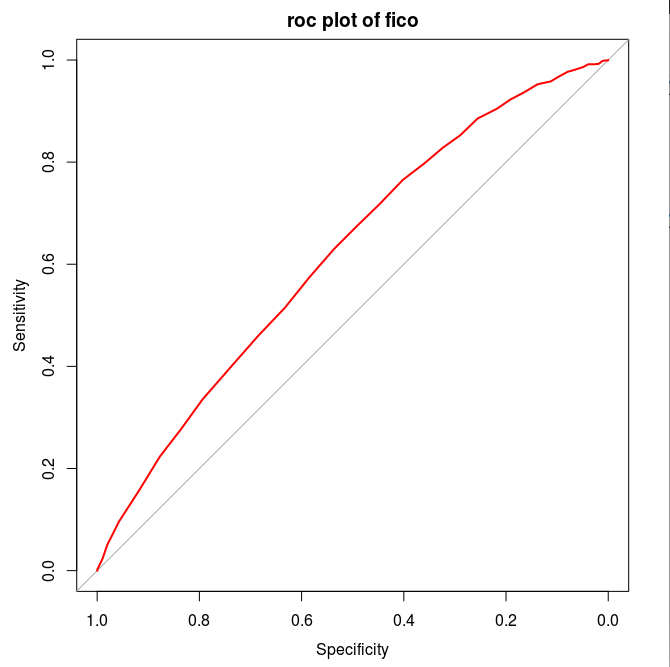
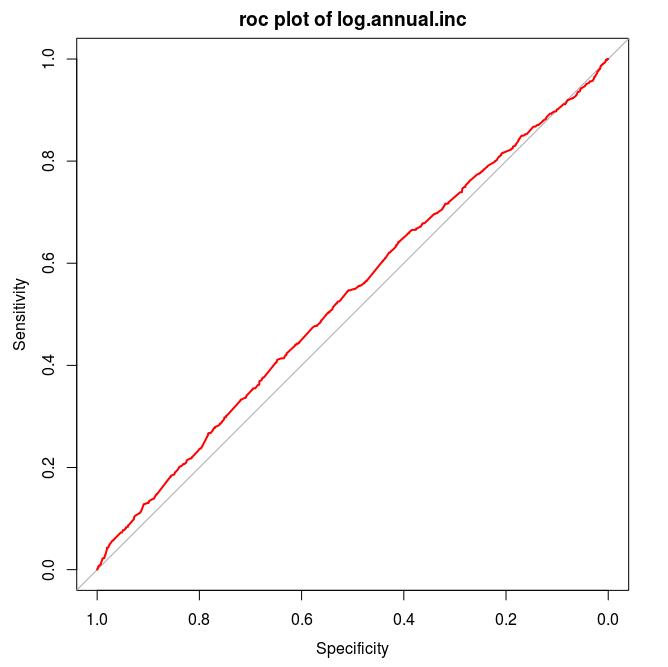
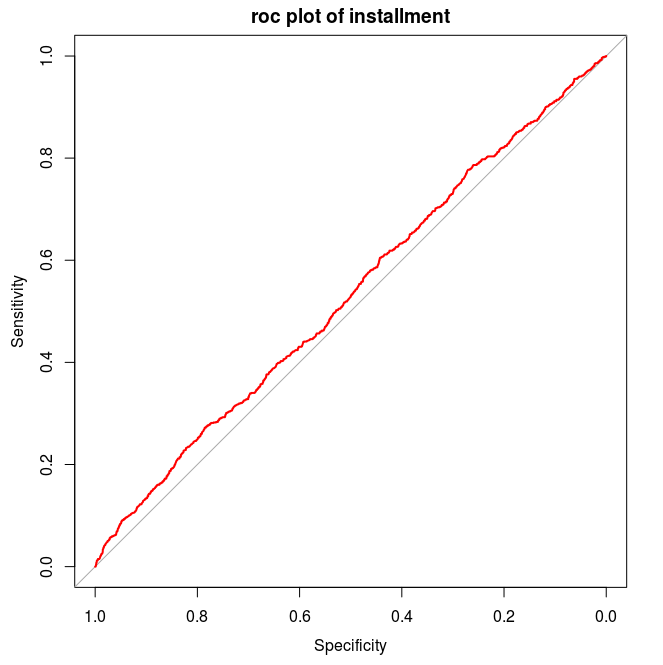
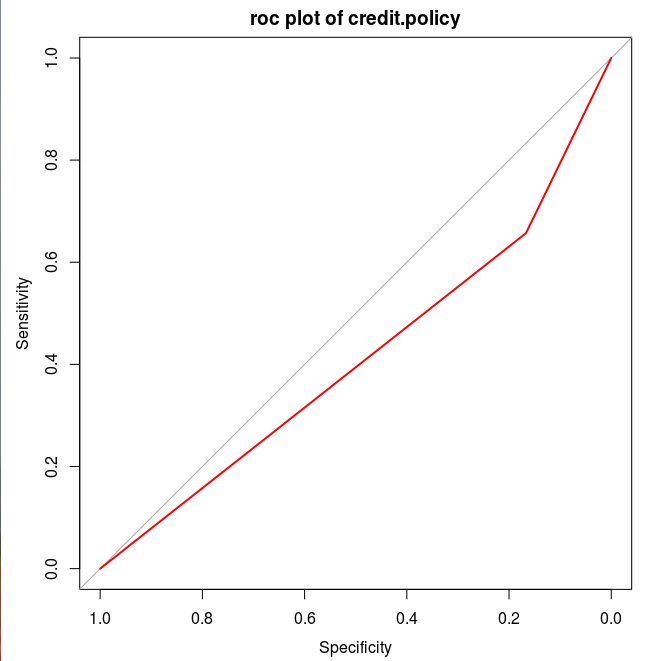
این ماتریس می‌تواند برای توصیف کارآیی[[4]](#footnote-3) مدل‌های دسته بندی مورد استفاده قرار گیرد. از این جدول برای تعیین مقدار شاخص های ارزیابی مانند دقت[[5]](#footnote-4) و صحت[[6]](#footnote-5) استفاده می شود. دقت عبارت است از اینکه "چه میزان از نمونه های انتخابی درست هستند" و صحت بر این مفهوم که "چه میزان از نمونه های صحیح موجود انتخاب شده اند" دلالت دارند.

برای بدست آوردن کارآیی یک دسته بندی کننده، کافی است مجموع عناصر قطر اصلی را بر مجموع کل عناصر ماتریس تقسیم نمود.

جدول زیر مثالی است، از ماتریس سردرگمی:



* **Accuracy**: Overall, how often is the classifier correct?
* (TP+TN)/total = (100+50)/165 = 0.91
* **Misclassification Rate**: Overall, how often is it wrong?
* (FP+FN)/total = (10+5)/165 = 0.09
* equivalent to 1 minus Accuracy
* also known as "Error Rate"
* **True Positive Rate**: When it's actually yes, how often does it predict yes?
* TP/actual yes = 100/105 = 0.95
* also known as "Sensitivity" or "Recall"
* **False Positive Rate**: When it's actually no, how often does it predict yes?
* FP/actual no = 10/60 = 0.17
* **Specificity**: When it's actually no, how often does it predict no?
* TN/actual no = 50/60 = 0.83
* equivalent to 1 minus False Positive Rate
* **Precision**: When it predicts yes, how often is it correct?
* TP/predicted yes = 100/110 = 0.91
* **Prevalence**: How often does the yes condition actually occur in our sample?
* actual yes/total = 105/165 = 0.64



Aria under curve = 0.6721337

برای ارزیابی و مقایسه ی مدل‌ها می توان از کتابخانه های lmtest، pscl، MKmisc،survey استفاده کرد.

تابع anova نیز در base برای مقایسه‌‌ی دو مدل قابل استفاده است.

می‌توان از تابع anova در base و یا تابع lrtest در کتابخانه‌ی lmtest برای تست ضریب احتمال درستی[[7]](#footnote-6) استفاده کرد.

برخلاف رگرسیون خطی، در اینجا امکان محاسبه ی R^2 وجود ندارد اما با استفاده از کتابخانه‌ی pscl می‌توان Pseudo R^2 را محاسبه کرد.

Hosmer-Lemeshow Test و Wald Test معیارهای دیگری هستند که به ترتیب در کتابخانه‌های MKmisc و survey قابل استفاده هستند.

1. multiple imputation by chained equations [↑](#footnote-ref-0)
2. methods(mice) [↑](#footnote-ref-1)
3. Confusion matrix [↑](#footnote-ref-2)
4. Performance [↑](#footnote-ref-3)
5. Precision [↑](#footnote-ref-4)
6. Recall [↑](#footnote-ref-5)
7. Likelihood Ratio Test [↑](#footnote-ref-6)