**به­­ نام خدا**

**فصل 1**

**پیش پردازش داده­ها**

**نام مجموعه داده**:داده های دودرصد مردم از پایگاه رفاه ایرانیان

توضیحی در مورد مجموعه:

این فایل 1048576 نفر از مردم است که به نوعی نمونه برداری شده که اطلاعات آماری (توزیعی) نمونه ایجاد شده منطبق بر اطلاعات آماری جامعه اولیه است. این امر سبب می‌شود که نتایج تحلیل‌های انجام شده بر روی نمونه آماری، قابلیت تعمیم به جامعه آماری اولیه را داشته باشد.

جهت حفظ حریم شخصی افراد داده هایی که مشخص کننده ی خود فرد هستند وجود ندارد.

**تعداد رکورد ها**:1048576

**تعداد متغیر ها**:47

**نام متغیر ها:**

1. شناسه فرد ( پیوسته )
2. شناسه سرپرست (پیوسته)
3. جنسیت (۱:مرد - ۲: زن) (دودویی)
4. کدپستی (۷ رقم ابتدایی) (پیوسته)
5. استان (اسمی)
6. شهرستان (اسمی)
7. آیا فرد ساکن شهر است؟ (دودویی)
8. گردش بستانکار حسابهای فرد در سال۹۵ (پیوسته)
9. گردش بدهکار حسابهای فرد در سال ۹۵ (پیوسته)
10. مانده ابتدای سال حسابهای فرد در سال ۹۵ (پیوسته)
11. مانده انتهای سال حسابهای فرد در سال ۹۵ (پیوسته)
12. مجموع سود حسابهای فرد در سال ۹۵ (پیوسته)
13. گردش بستانکار حسابهای فرد در سال ۹۶ (پیوسته)
14. گردش بدهکار حسابهای فرد در سال ۹۶ (پیوسته)
15. مانده ابتدای سال حسابهای فرد در سال ۹۶ (پیوسته)
16. مانده انتهای سال حسابهای فرد در سال ۹۶ (پیوسته)
17. مجموع سود حسابهای فرد در سال ۹۶ (پیوسته)
18. گردش بستانکار حسابهای فرد در سال ۹۷ (پیوسته)
19. گردش بدهکار حسابهای فرد در سال ۹۷ (پیوسته)
20. مانده ابتدای سال حسابهای فرد در سال ۹۷ (پیوسته)
21. مانده انتهای سال حسابهای فرد در سال ۹۷ (پیوسته)
22. مجموع سود حسابهای فرد در سال ۹۷ (پیوسته)
23. گردش بستانکار حسابهای فرد در سال ۹۸ (پیوسته)
24. گردش بدهکار حسابهای فرد در سال ۹۸ (پیوسته)
25. مانده ابتدای سال حسابهای فرد در سال ۹۸ (پیوسته)
26. مانده انتهای سال حسابهای فرد در سال ۹۸ (پیوسته)
27. مجموع سود حسابهای فرد در سال ۹۸ (پیوسته)
28. مقدار تراکنش کارتهای فرد در سال 98 (پیوسته)
29. تعداد تراکنش کارتهای فرد در سال 98 (پیوسته)
30. مقدار تراکنش کارتهای فرد در۶ ماه ابتدای سال ۹۹ (پیوسته)
31. تعداد تراکنش کارتهای فرد در ۶ ماه ابتدای سال ۹۹ (پیوسته)
32. آیا فرد بیمارخاص است؟ (دودویی)
33. آیا فرد معلول است؟ (دودویی)
34. آیا فرد بیمه‌پرداز صندوقهای بازنشستگی است؟ (دودویی)
35. آیا فرد بازنشسته صندوقهای بازنشستگی است؟ (دودویی)
36. آیا فرد شاغل مشمول مالیات است؟ (دودویی)
37. مجموع درامد فرد (از حقوق) (پیوسته)
38. تعداد خودروهای فرد (پیوسته)
39. مجموع ارزش خودروهای فرد (پیوسته)
40. تعداد سفرهای خارجی هوایی غیرزیارتی ۹۶تا۹۹ (پیوسته)
41. تعداد سفرهای خارجی زمینی غیرزیارتی ۹۶تا۹۹ (پیوسته)
42. تعداد سفرهای خارجی هوایی زیارتی ۹۶تا۹۹ (پیوسته)
43. تعداد سفرهای خارجی زمینی زیارتی ۹۶تا۹۹ (پیوسته)
44. آیا فرد دارای مجوز صنفی است (دودویی)
45. مجوز صنفی (اسمی)
46. آیا فرد دارای بیمه سلامت است (دودویی)
47. نوع بیمه سلامت (اسمی)

**هدف:**

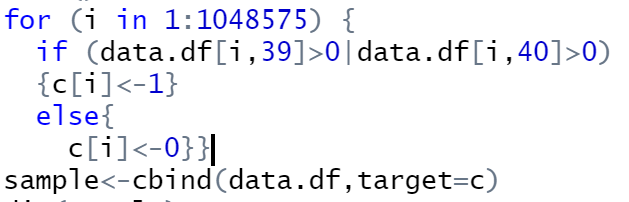
رده بندی افراد بر اساس رفتن یا نرفتن به سفر خارجی غیر زیارتی

**ضرورت هدف:**

یک آژانس مسافرتی برای بلیغاتش نیاز دارد بداند که تبلیغاتش را برای چه افرادی بفرستد بازدهی بالاتری دارد و آنها تررغیب می­شوند به سفر بروند و درنتیجه اینکار با هزینه کمتر بهترین نتیجه را از تبلیغاتش خواهد­گرفت.

**توضیحاتی در مورد متغیرها:**

در داده های موجود متغیر برآمد که تعیین سفر خارجی رفتن یا نرفتن بود را نداشتیم بنابراین با نوشتن کد آن در R این متغیردودویی(0و1) را با استفاده از دو متغیر" تعداد سفرهای خارجی هوایی غیرزیارتی ۹۶تا۹۹" و " تعداد سفرهای خارجی زمینی غیرزیارتی" نیز به متغیر­هایمان افزودیم و آن دو متغیر استفاده شده را حذف می­کنیم.

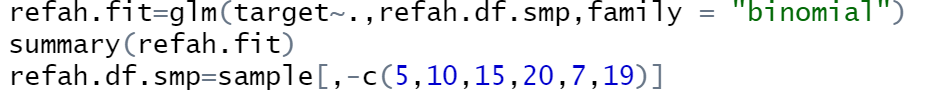


متغیرهای "استان" ، "شهرستان" ، "مجوز­صنفی" ، "نوع بیمه سلامت" به دلیل مخدوش بودن در فایل گرفته شده حذف گردیدند

متغیر های "شناسه فرد" ،"شناسه سرپرست" ،"کدپستی" اطلاعات مفیدی برای تعیین برآمد به ما نمی­دهند؛ بنابراین حذف گردیدند.

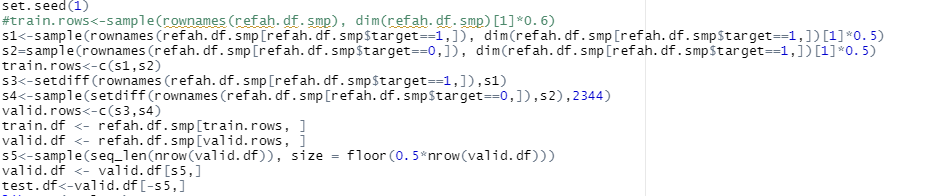
حال باتوجه به متغیرها و کارایی آن­ها متوجه شدیم که با متغیر "گردش بستانکار حسابهای فرد" ، "گردش بدهکار حسابهای فرد" ،"مانده انتهای سال حسابهای فرد " درهر یک از سال های­ 95،96،97و98 می­توانیم متغیر"مانده ابتدای سال حسابهای فرد" در همان سال را به دست آوریم بنابراین این 4متغیر نیز حذف میگردند.

پس از آن با استفاده از کد مدل رگرسیون لجستیک p-value های هرمتغیر نسبت به متغیر برآمد را محاسبه می­کنیم و ازین جا متوجه می­شویم که دو متغیر "مجموع سود حسابهای فرد در سال ۹۵" و"گردش بدهکار حسابهای فرد در سال ۹۸" و" اثری بر روی برآمد نداشته و می­توان آنهارا نیز حذف کرد.



تفکیک داده های به مجموعه داد­ه­های آموزشی و اعتبارسنجی و آزمون

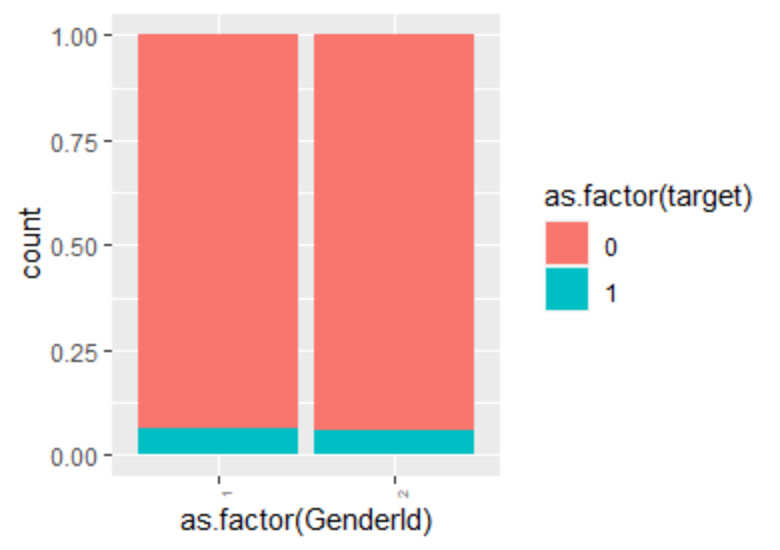
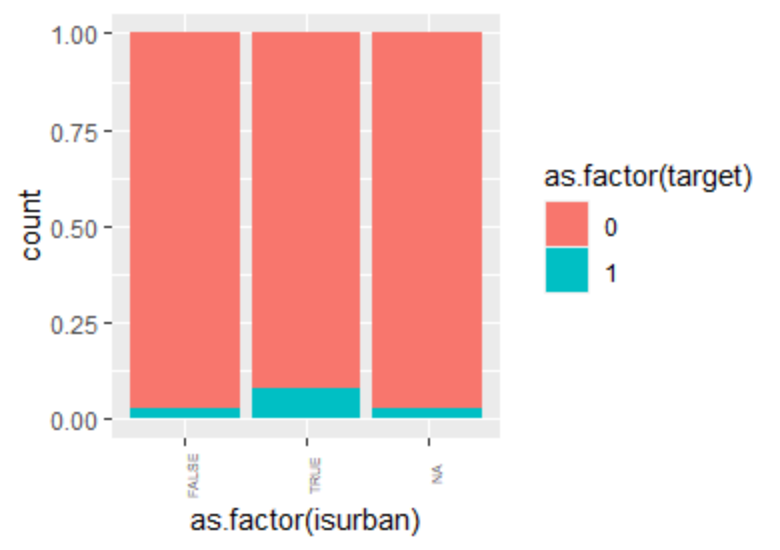
باتوجه به آنکه تعداد 0 و1 ها در مجموعه داده بسیار نامتوازن هستند اگر به صورت عادی و رندوم به دسته بندی شوند مدل هایمان فریب خورده و تواناییشان برای تشخیص درست1 ها بسیارکم می­شود؛ بنابراین از روش بیش­نمونه­گیری(oversampling) برای بخش بندی داده­ها به داده­های آموزشی و اعتبار­سنجی استفاده شد که در ادامه کد آن را مشاهده می­شود.



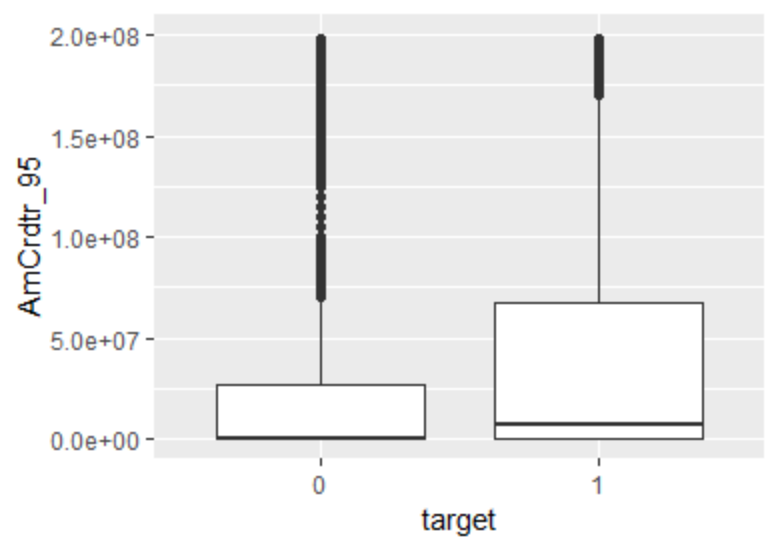
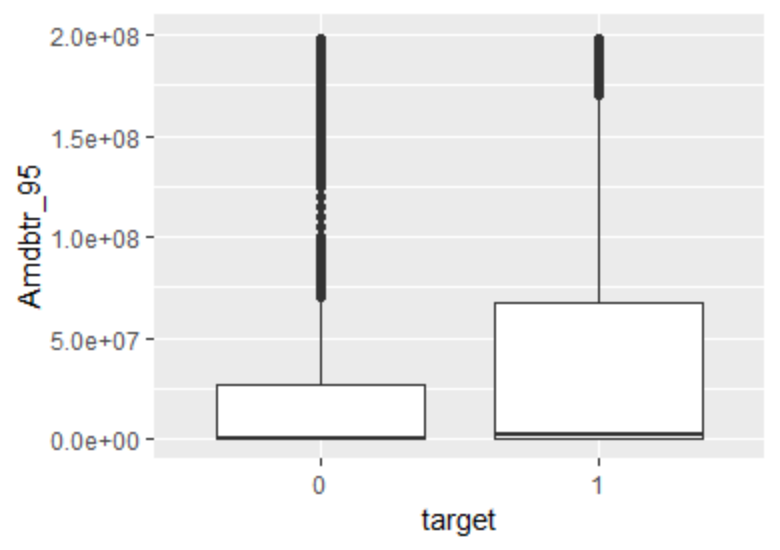
**فصل 2**

**تصویری سازی**

در این فصل به بررسی نمودار تمام متغیر ها نسبت به برآمد پرداختیم تا متوجه اثر بخشی یا بی اثر بودن هر متغیر بر روی متغیر برآمد بشویم.

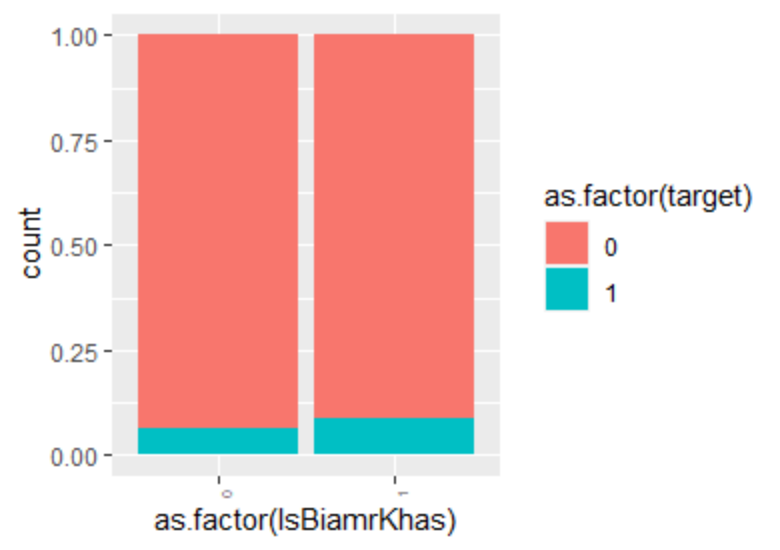
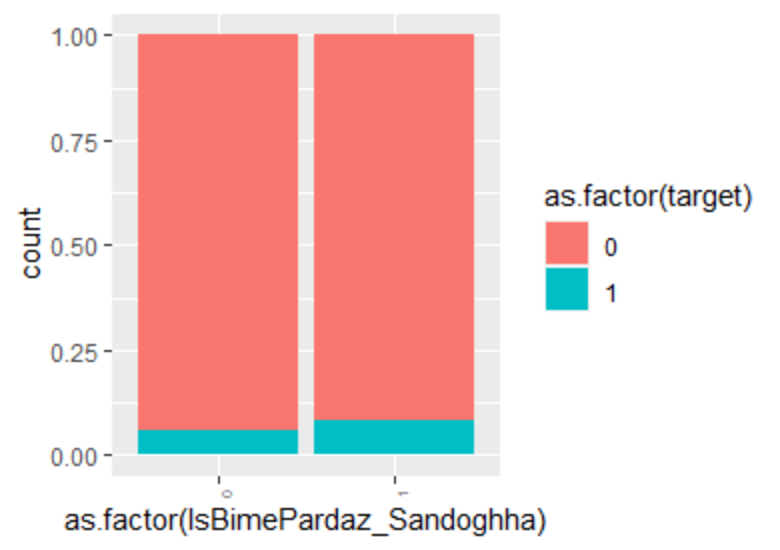


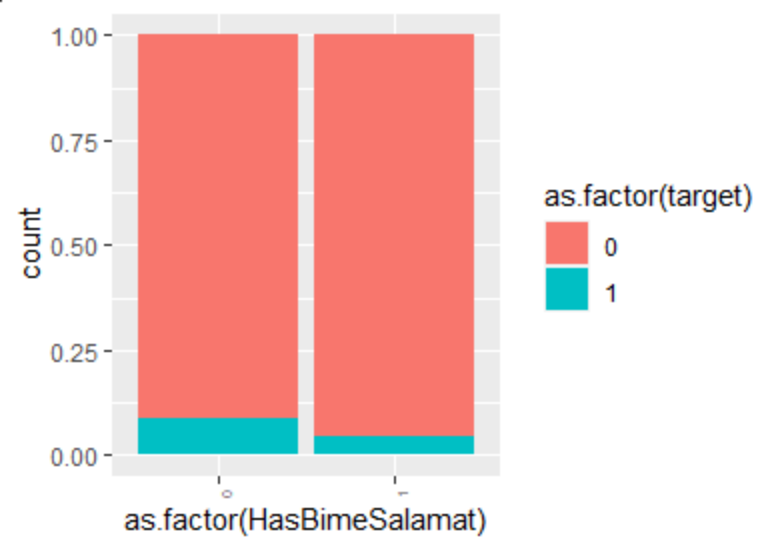
همانگونه که در نمودارهای فوق پیداست اثر دومتغیر جنسیت و ساکن شهر یا روستا بودن بر روی متغیر برآمد بسیار اندک است.



همانطور که انتظار میرفت رفتار این دو نمودار گردش بستانکارهای حساب فرد گردش بدهکارهای حساب فرد رفتاری مشابه یکدیگر دارند.

این اتفاق در سال های 96و97و98 نیز رخ می­دهد.

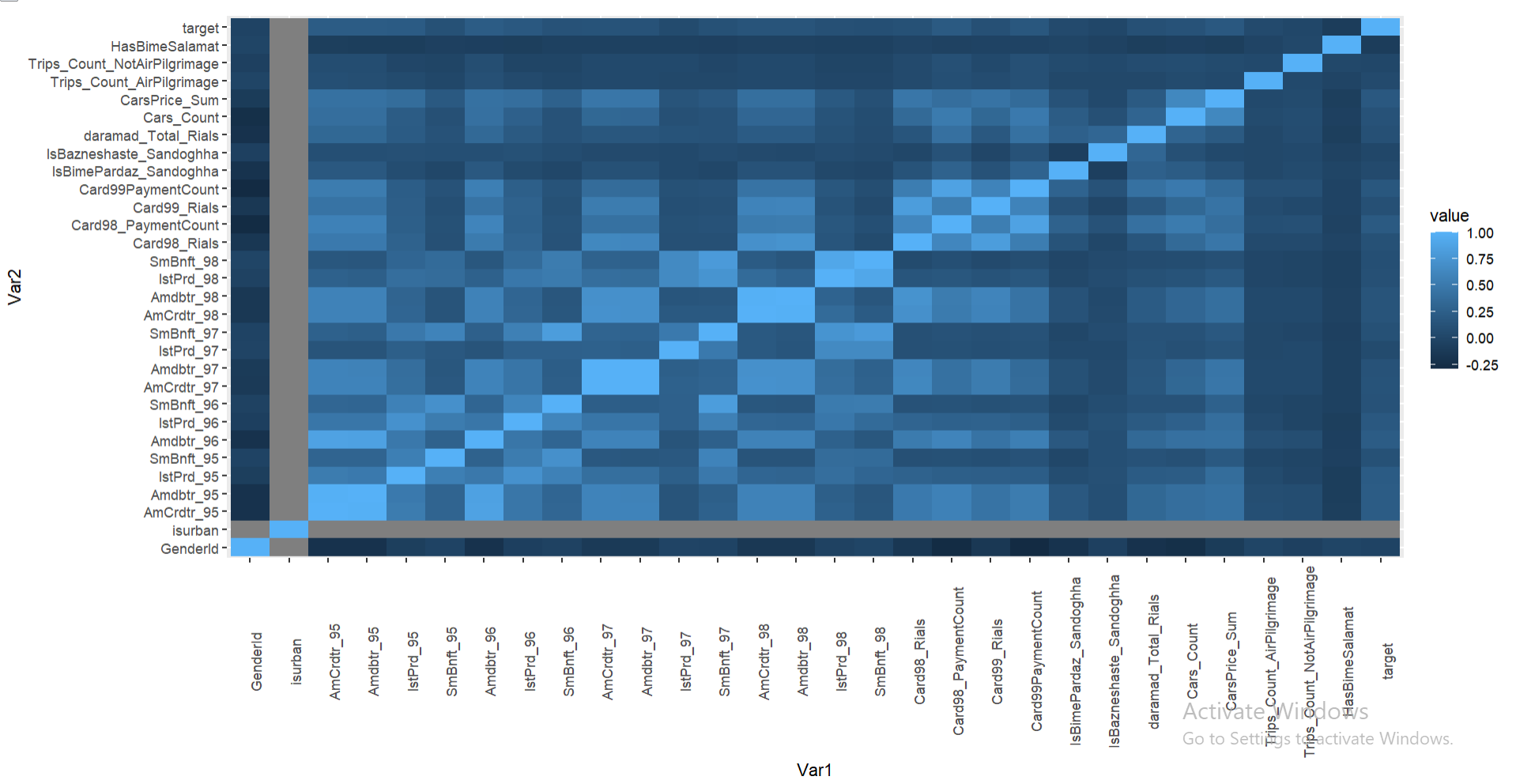




باتوجه به سه نمودار فوق متوجه اثرکم این سه متغیر بر روی برآمد میشویم.

سایر متغیر ها برروی داده مان اثربخش بوده و به همین جهت در این گزارش آورده نشده اند.

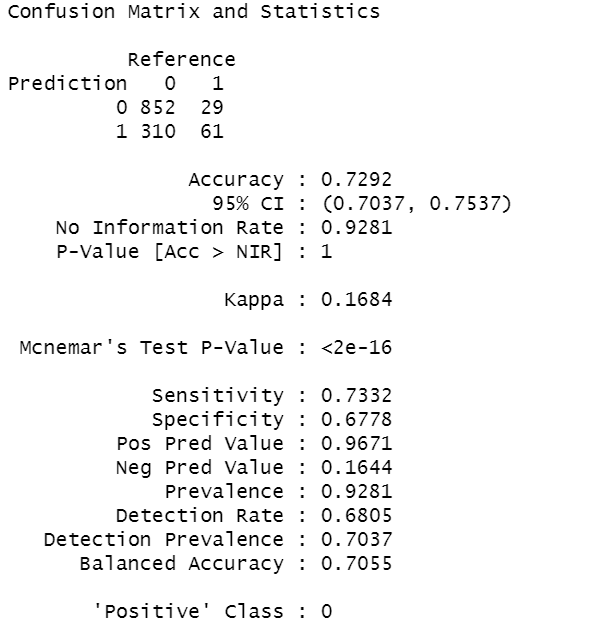
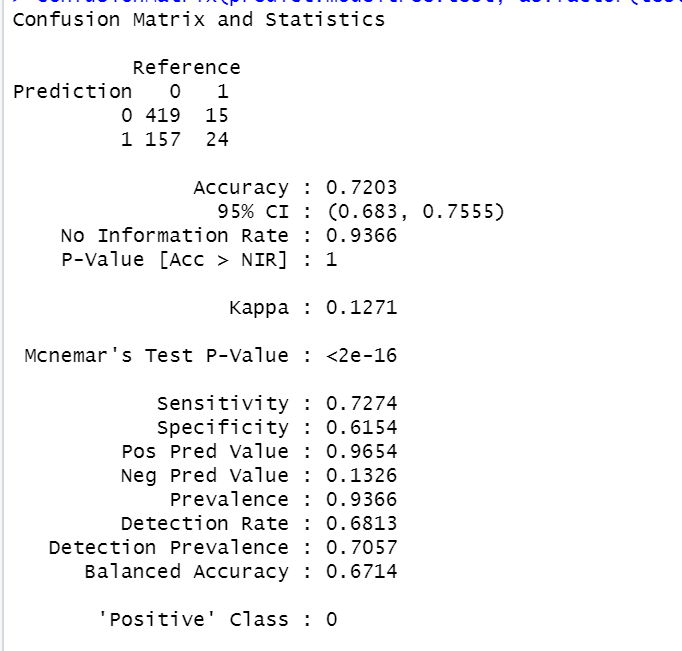
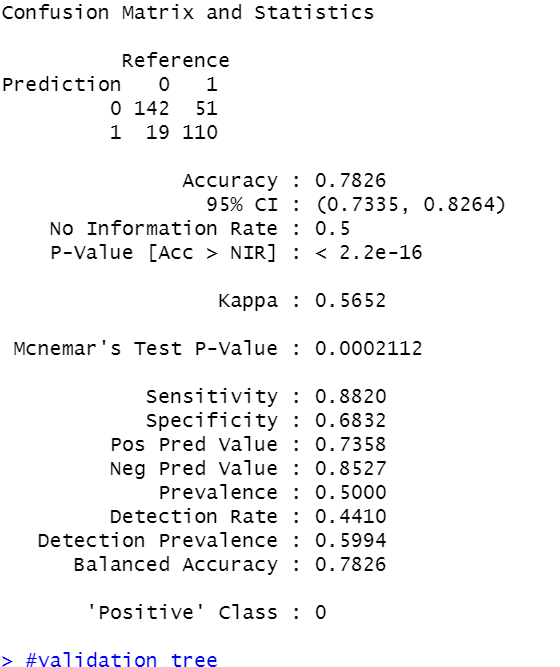
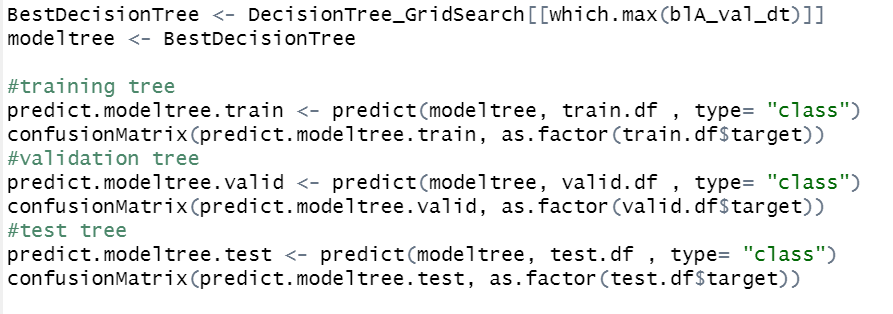
با نگاهی به نمودار حرارتی زیر صحه بر صحبت های قبل گذاشته می­شود و وابستگی گردش حساب­های بدهکار وبستانکار فرد را نمایش می­دهد.



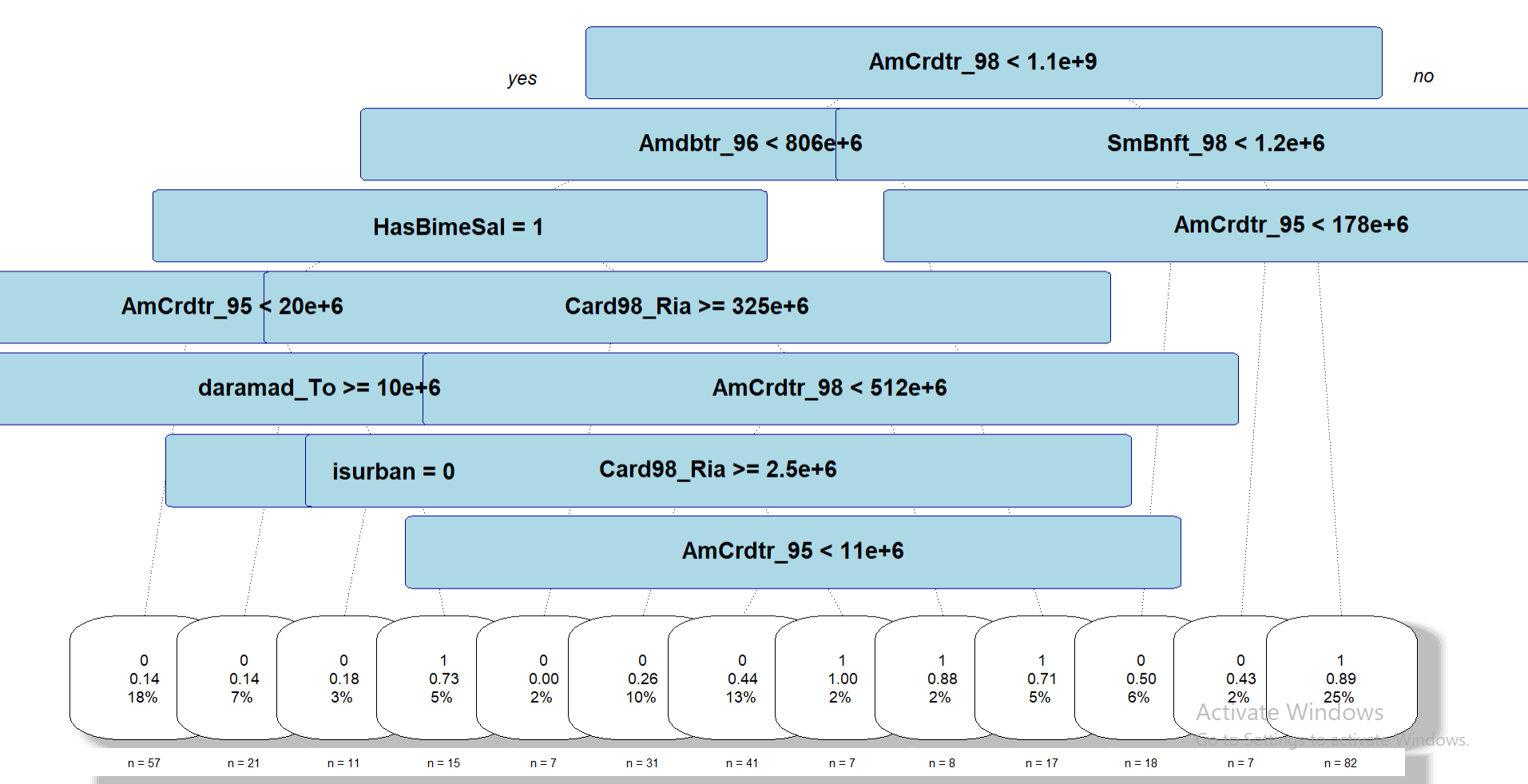
**فصل 3**

**برازش مدل درخت**

مدل درخت را روی مجموعه داده­ی آموزشی و سپس مجموعه داده­ی با استفاده از گریدسرچ اعتبار سنجی برازش داده شد که نتیجه را در زیر مشاهده می­شود:

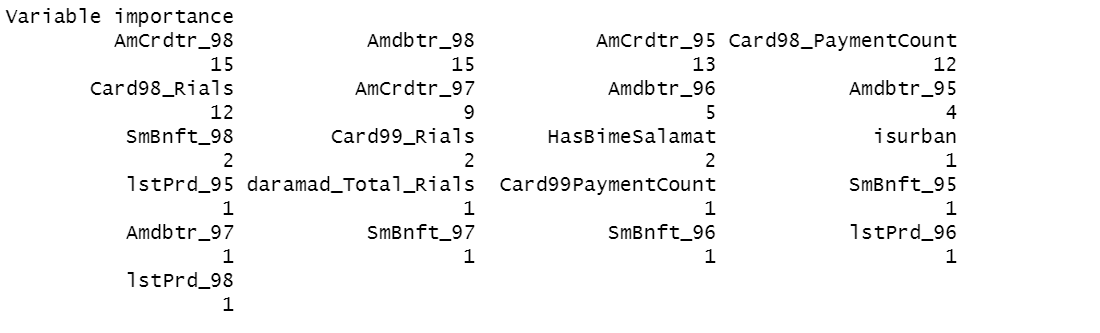


تصویر بالا سمت چپ نتایج به­دست آمده از برازش مدل بر روی مجموعه آموزشی وتصویر پایین سمت چپ نتایج به دست آمده بر روی مجموعه ی اعتبار سنجی ونتایح به دست آمده در تصویر پایین سمت راست مربوط به برازش مدل بر روی مجموعه ی آزمون است که به ما می­گوید دقت این مدل 72 درصد می­باشد.



نمودار درخت والگوریتم آن نیز براساس شکل بالا می­باشد.

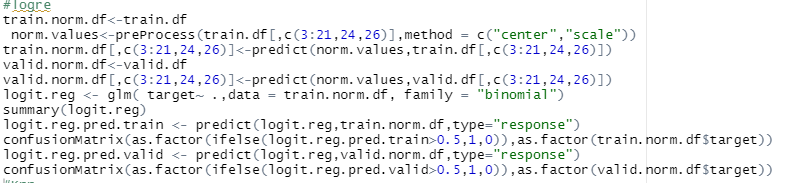
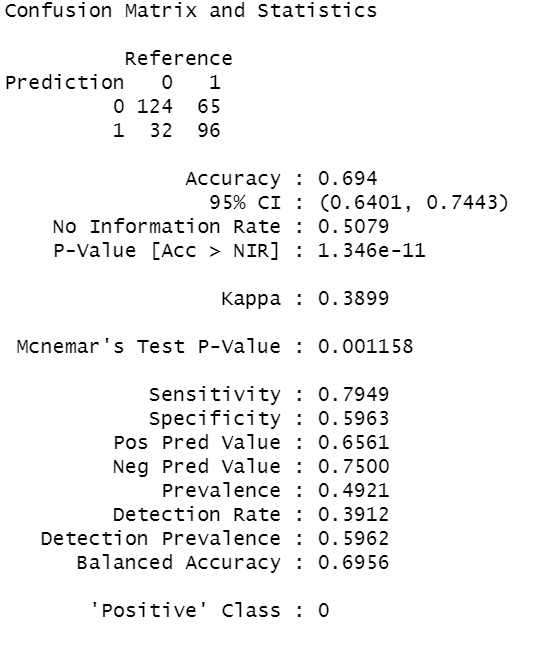
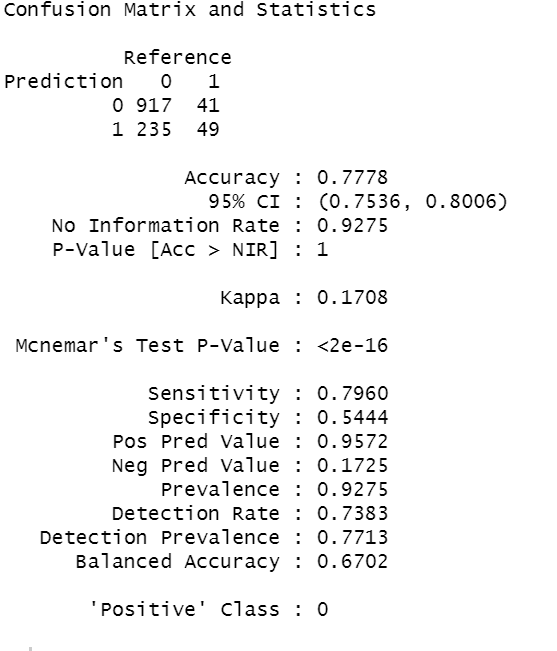
با summary گرفتن از مدل به متغیرهای پراهمیت تر در مدل دست می­یابیم:



**فصل4**

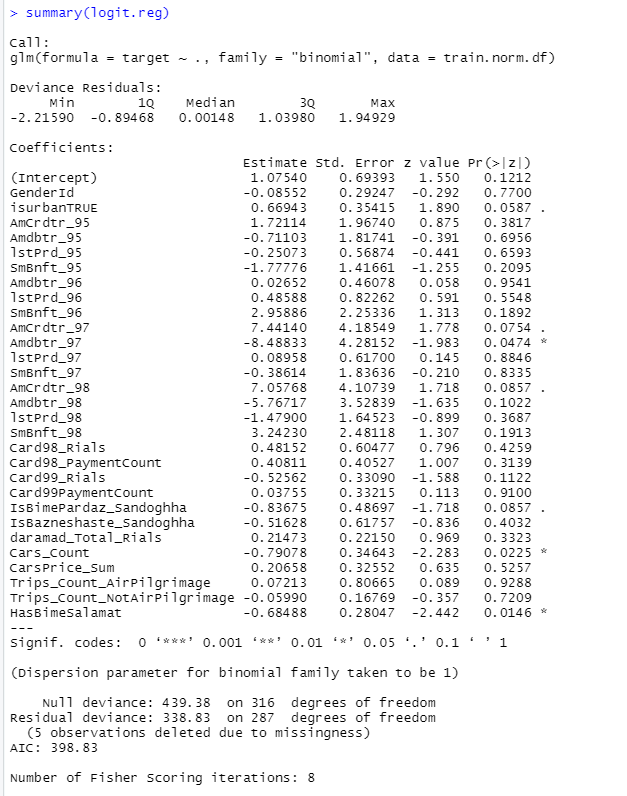
**برازش مدل رگرسیون لجستیک**

مدل رگرسیون لجستیک را بر روی داده­های آموزشی و اعتبارسنجی می­برازانیم و نتیجه را مشاهده می­کنیم:

تصویر سمت راست دقت حاصل شده از برازش مدل بر روی داده های آموزشی و تصویر سمت چپ بر روی داده های اعتبارسنجی است و بنابرنتایج به دست آمده دقت این مدل 77% می باشد.

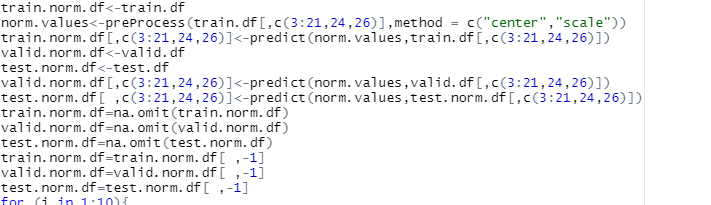
و با summary گرفتن از مدلمان نتایج زیر و متغیر های مفید مشاهده می­شوند.



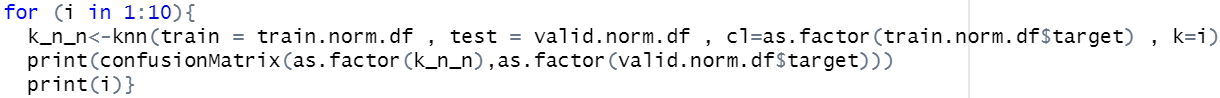
**فصل 5**

**برازش مدل knn**

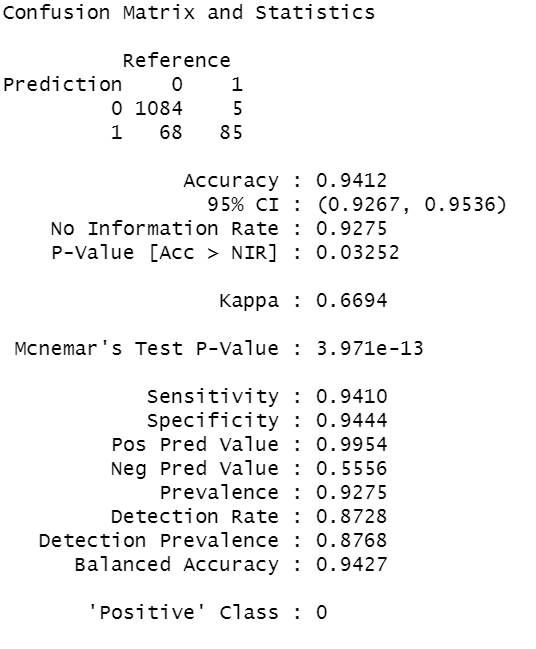
مجموعه ها را نرمال سازی کرده و داده های گمشده را خارج شده است.



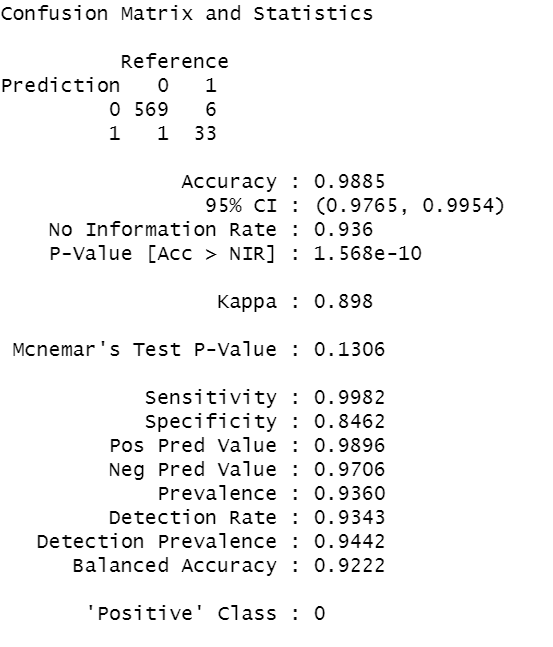
سپس مدل را بر روی مجموعه آموزشی برازانده و بهترین k را با استفاده از مجموعه اعتبارسنجی یافت شد.



6= k بهترین kمی باشد که نتیجه اش به شکل زیراست.



حال با 20=k نتیجه را برروی داده­آزمون مشاهده می­شود:

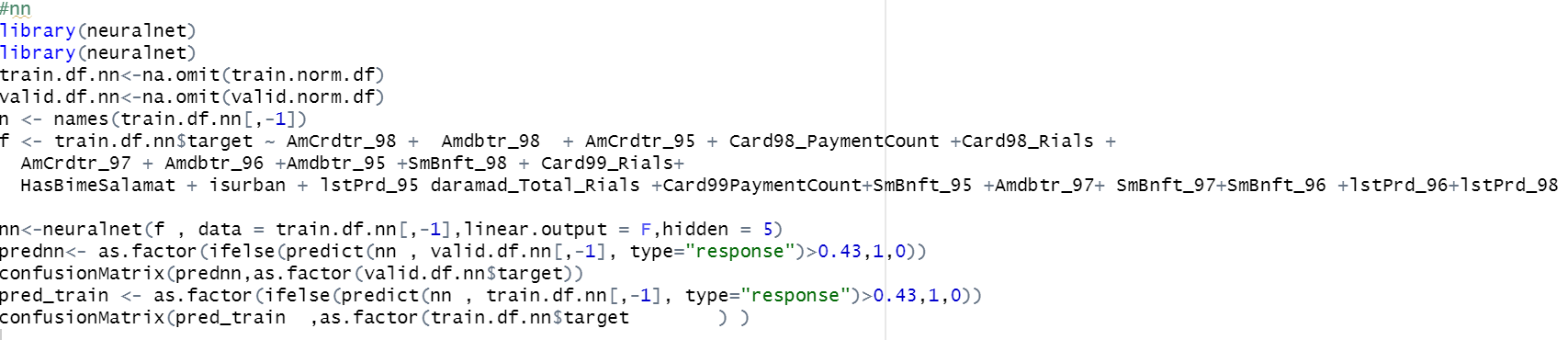


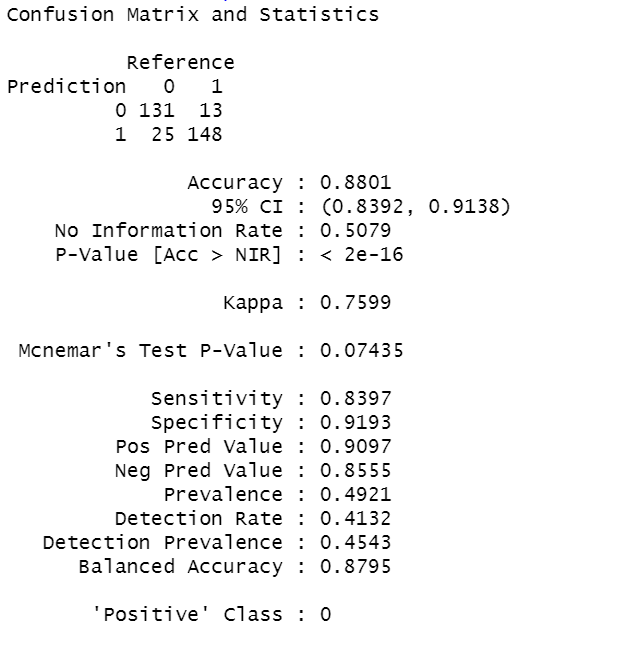
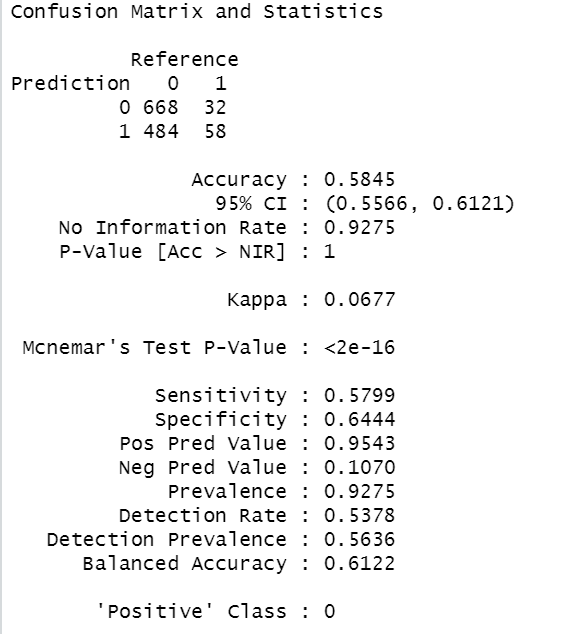
همانطور که مشاهده می­ شود دقت این مدل98% می­باشد.

**فصل 6**

**برازش مدل شبکه عصبی(nn)**

با اجتماع گرفتن از متغیر های خوبی که در خلاصه ی مدل های درخت و رگرسیون یافتیم نوعی فیچرسلکشن انجام شده و و تنها آن ها را وارد مدل­ می­شوند وبرازش داده می­شوند.



تصویرسمت راست مربوط به مجموعه آموزشی و تصویر سمت چپ مربوط به مجموعه اعتبار سنجی می­باشد و نشان دهنده آن است مدل دچار بیش برازش شده است.

**فصل7**

**نتیجه گیری**

|  |  |
| --- | --- |
| **دقت** | **مدل** |
| 72% | **درخت** |
| 77% | **رگرسیون لجستیک** |
| 98% | **knn** |
| بیش برازش | **nn** |

با توجه به جدول بالا مدل knn در این پروژه بهترین مدل است.