بسم الله الرحمن الرحیم

پروژه ی درس داده کاوی

استاد راهنما: دکتر محمد رضا فقیهی حبیب آبادی

دانشجو: احمد یزدانی

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

تیرماه 1401

مجموعه داده های هزینه و درآمد خانوارهای شهری 1399

این مجموعه داده از شهرهای هرمزگان، خوزستان،بوشهر و فارس جمع آوری شده اند.

تعداد ثبت ها برابر با 2656 است.

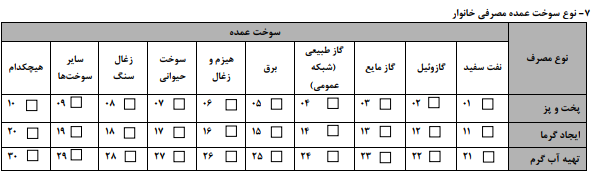
تعداد متغییر ها 67تا است.

متغییر ها عبارتند از:

1. نام استان: متغییری رسته ای که شامل خوزستان، هرمزگان و فارس است.
2. آدرس: آدرس خانوار که با یک عدد یازده رقمی مشخص میشود.
3. نوع خانوار: یک متغییر دودویی که نشان دهنده ی سکونت معمولی ساکن یا گروهی است.
4. تعداداعضا: متغییری گسسته که نمایش دهنده ی تعداد اعضای خانوار است.
5. جنسیت سرپرست: متغییری دودویی که نشان دهنده ی جنسیت سرپرست است.
6. سن سرپرست: متغییری گسسته که نشان دهنده ی سن سرپرست است.
7. سواد سرپرست: متغییری دودویی که نشان دهنده ی باسواد بودن سرپرست است.
8. تحصیل سرپرست:متغییری دودویی که نشان دهنده ی درحال تحصیل بودن یا نبودن سرپرست است.
9. مدرک سرپرست: متغییری رسته ای از 1 تا 9 که نشان دهنده سطح مدرک سرپرست است.
10. فعالیت سرپرست: متغییری رسته ای از عدد 1 تا 6 که نشان دهنده 1. شاغل 2.جویای کار 3. دارای درآمد بدون کار 4.محصل 5.خانه دار 6.سایر
11. زناشویی سرپرست: متغییری رسته ای از عدد 1 تا 4 که نشان دهنده 1.همسردارد 2.ندارد فوت 3. ندارد طلاق 4.ازدواج نکرده
12. نحوه تصرف: متغییری رسته ای از عدد 1 تا 7 که نشان دهنده 1.ملکی عرصه و اعیان 2.ملکی اعیان 3. اجاری 4.رهن 5.دربرابرخدمت 6.رایگان 7.سایر
13. تعداد اتاق: متغییری گسسته که نشان دهنده ی تعداد اتاق های خانه است.
14. سطح زیربنا: متغییری گسسته که نشان دهنده ی متراژ زیر بنای خانه است.
15. نوع اسکلت: متغییری رسته ای از اعداد 1 تا 3 که نشان دهنده ی 1.فلزی 2.بتون 3.سایر
16. ماشین: متغییری رسته ای از 1 و 0 که نشان دهنده داشتن یا نداشتن ماشین است.
17. موتور: متغییری رسته ای از 1 و 0 که نشان دهنده داشتن یا نداشتن موتور است.

از متغییر 18 تا 51 نشان دهنده داشتن و یا نداشتن وسایلی مانند دوچرخه، رادیو، رادیو ضبط، تلویزیون، تلویزیون رنگی و .... تا فاضلاب شهری است.

ستون های 52 و 53 و 54 نشان دهنده نوع سوخت مصرفی با توجه به شکل زیر است:



ستون های 55 تا 66 متغییر های پیوسته که به ترتیب نشان دهنده ی هزینه خوراکی، هزینه ی نوشیدنی، هزینه ی پوشک، هزینه ی مسکن، هزینه ی لوازم خانگی، هزینه ی درمانی، هزینه ی حمل ونقل، هزینه ی ارتباطات، هزینه ی تفریحات فرهنگی، هزینه ی غذا آماده، هزینه ی کالامتفرقه، هزینه ی کالا بادوام است.

ستون 67 نشان دهنده ی درآمد خانوار است.

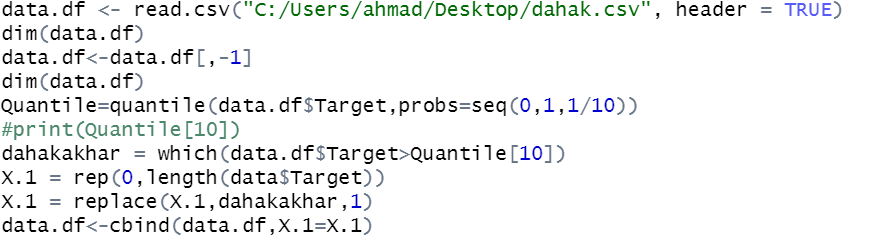
هدف

باتوجه به ستون درآمد میخواهیم دهک دهم را از سایر دهک ها جدا کنیم.

ضرورت هدف

باتوجه به سیاست های دولت، با جدا سازی دهک دهم از سایر دهک ها به سایر دهک ها یارانه تعلق میگیرد ولی به دهک دهم یارانه تعلق نمیگیرد.

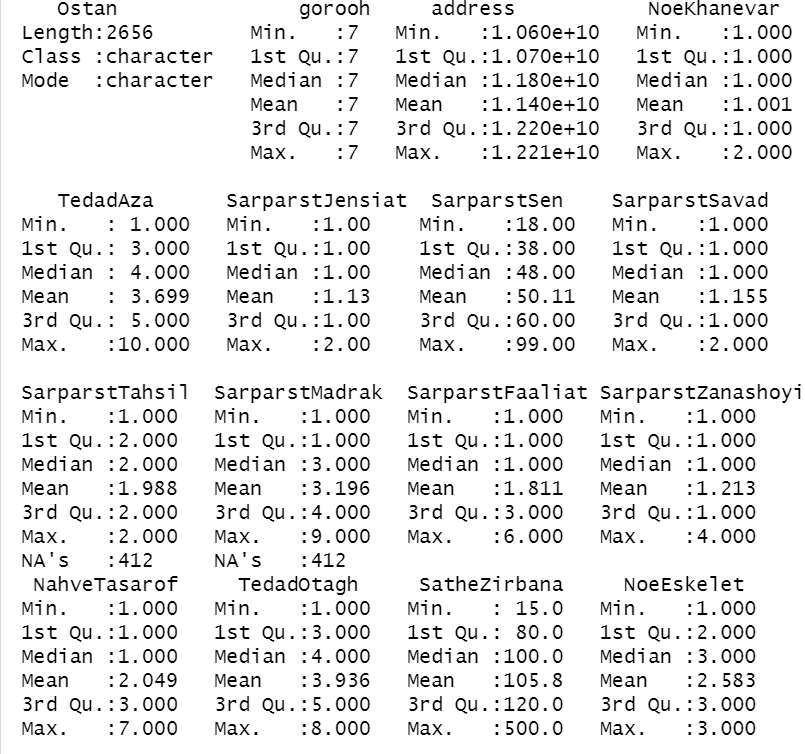
ابتدا متغییری کمکی به صورت دودویی از روی متغییر درآمد میسازیم به طوری که اگر متغییر درآمد بیشتر از 94075000 بود متغییر برآمد 1 شود در غیر اینصورت 0 شود.

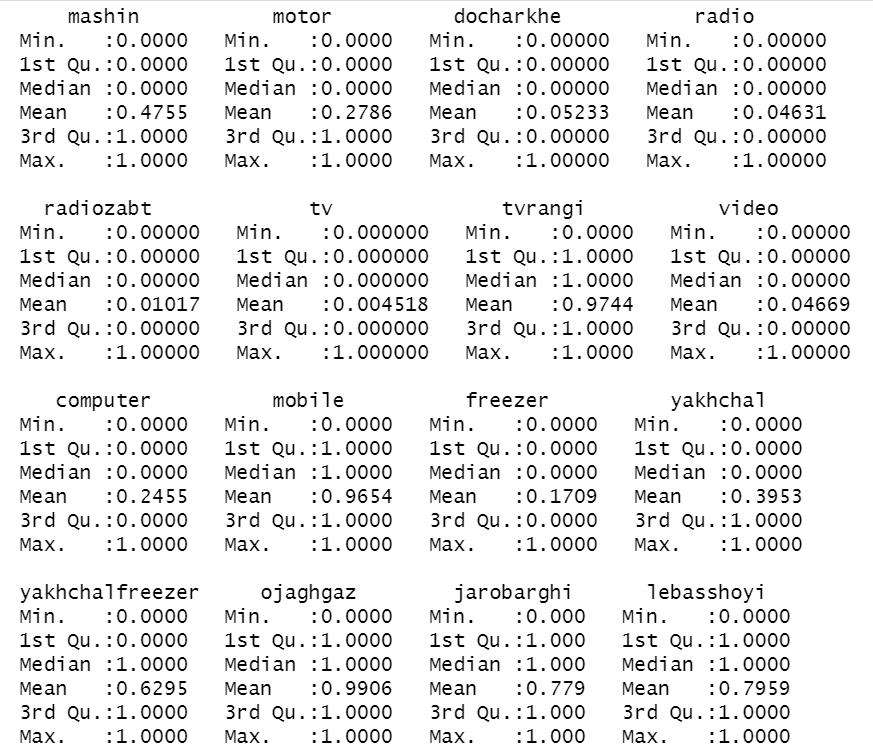


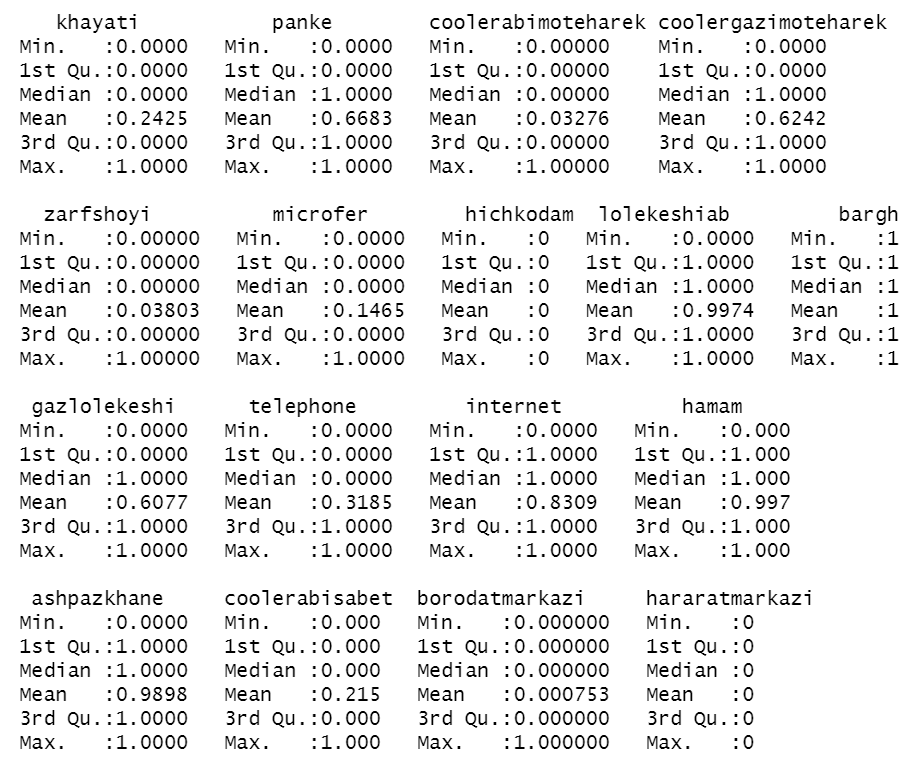
تشخیص 1 ها برای ما اهمیت بیشتری دارد زیرا اگر به شخصی اشتباه یارانه داده نشود آن شخص اعتراض میکند ولی اگر به شخصی اشتباه یارانه داده شود این اتفاق نمی افتد.

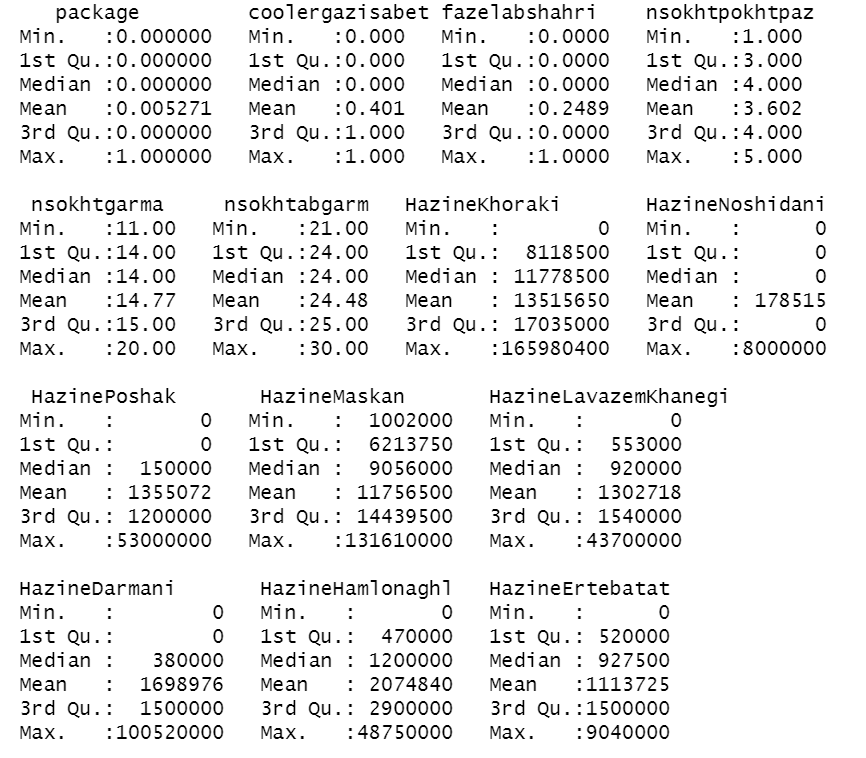
از طرفی تعداد 0 متغییر برآمد ها نه برابر تعداد 1 های متغییر برآمد است پس 1 های متغییر برآمد ارزش بیشتری برای ما دارند.

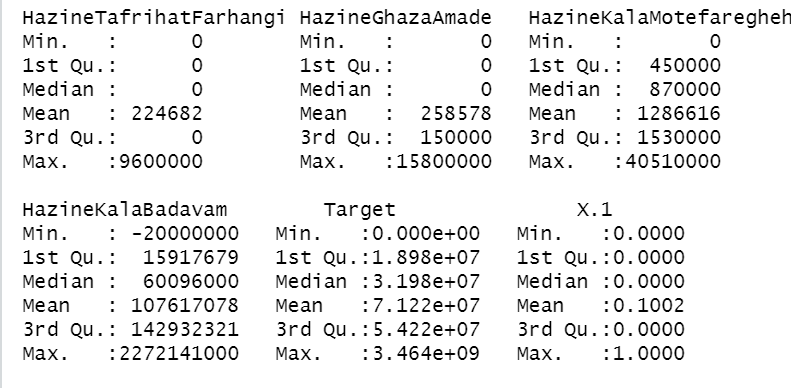
حال خلاصه ای از داده هارا در زیر نشان میدهیم:









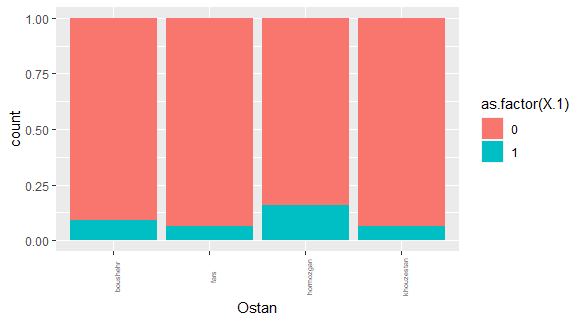


متغییر های آدرس، گروه، برق، هیچکدام و حرارت مرکزی را حذف مینماییم.

زیرا متغییر آدرس اطلاعات خاصی به ما نمیدهد و متغییر برق و هیچکدام و حرارت مرکزی در تمام ثبت ها یک مقدار ثابت در بر میگیرند.

همچنین در ستون تحصیل سرپرست و مدرک سرپرست مقادیر گمشده را حذف مینماییم.

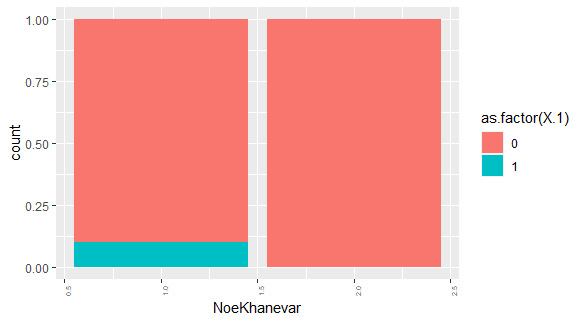
حال با رسم نمودار متغییر های پیش گو برحسب هدف داریم:



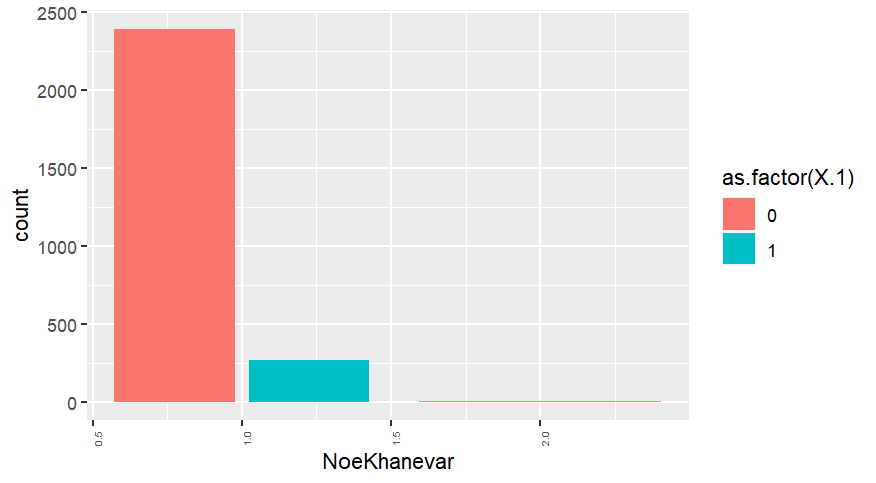
با توجه به نمودار اکثریت 1 ها در هرمزگان و کمترین 1 ها در استان فارس قرار دارند.

551 نفر استان بوشهر، 654 استان فارس، 860 نفر استان هرمزگان و 591 نفر استان خوزستان.

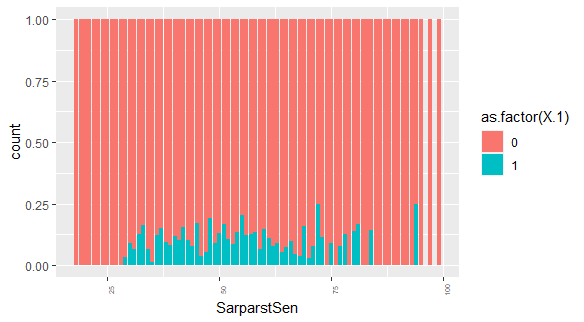
با رسم نمودار نوع خانوار برحسب متغییر برآمد داریم:

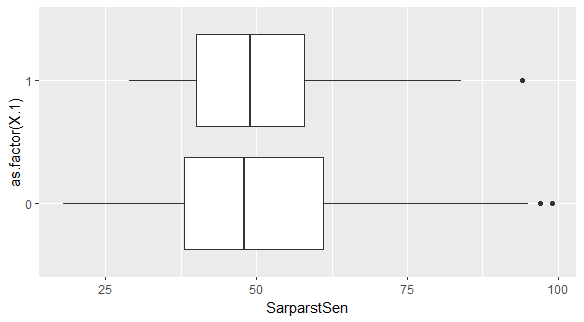


به ظاهر با توجه به نمودار اگر نوع خانوار در رده ی 2 قرار بگیرد نمیتواند متغییر برآمد در رده ی 1 قرار بگیرد.

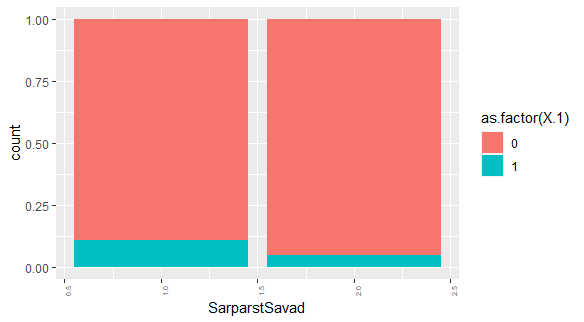


حال با بررسی مجدد مشاهده میشود تعداد خانوار نوع 2 بسیار کم است(فقط 2 خانوار) پس این پیش گو میتواند کاندید حذف باشد.

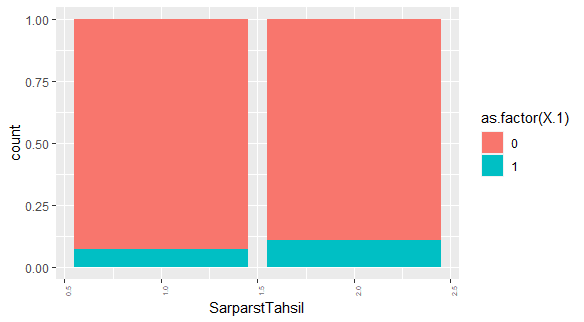




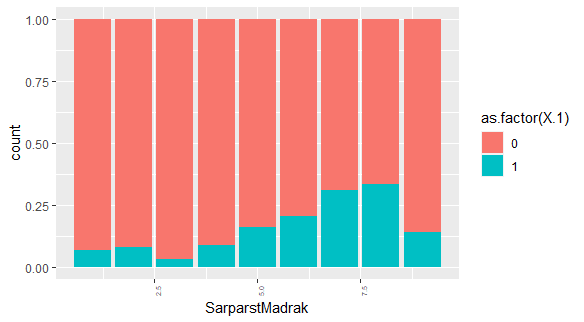
با توجه به نمودار سن سرپرست اکثریت 1های برآمد در سن های 30 تا 80 اتفاق میفتد.



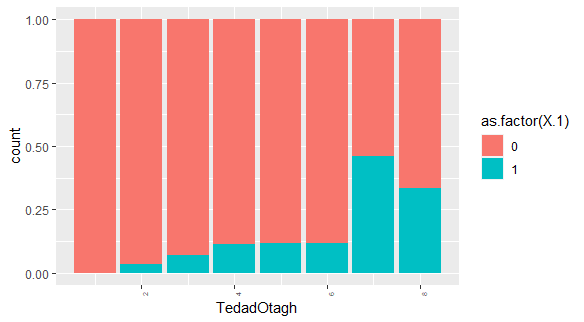
باتوجه به سواد سرپرست، اگر سرپرست سواد داشته باشد احتمال بیشتری متغییر برآمد 1 میشود.



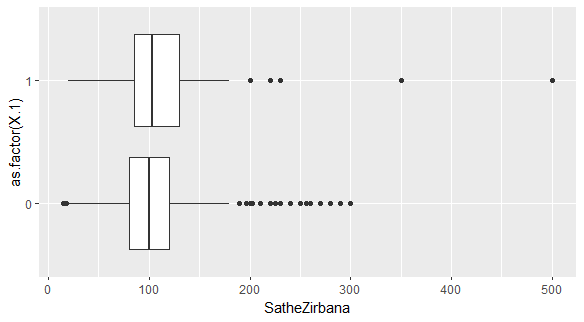
با توجه به نمودار فوق اگر سرپرست درحال تحصیل نباشد احتمال بیشتری متغییر برآمد 1 است. در این ستون 412 ثبت گمشده وجود دارد.(تعداد 1 های این پیش گو 27 تا است )



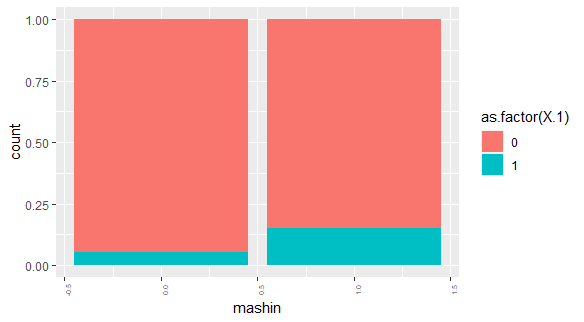
باتوجه به نمودار فوق هرچه مدرک تحصیلی سرپرست بالاتر باشد احتمال 1 شدن متغییر برآمد بیشتر است. در این ستون 412 ثبت گمشده وجود دارد.



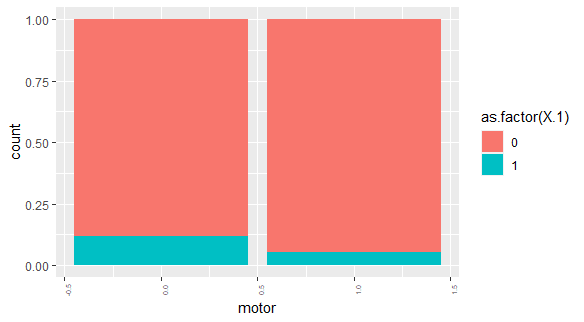
باتوجه به نمودار فوق هرچه تعداد اتاق ها بیشتر باشند احتمال بیشتری متغییر برآمد 1 خواهد شد.



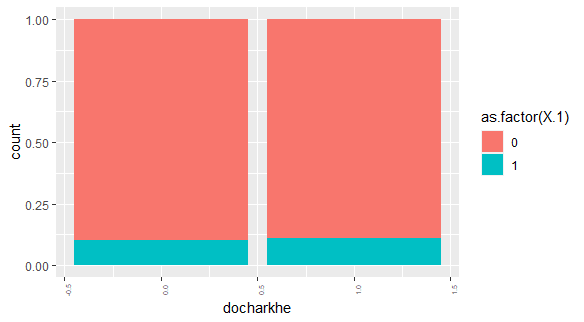
باتوجه به نمودار جعبه ای سطح زیر بنا در میابیم که در جاهایی که متغییر برآمد 1 است سطح زیربنای بیشتری دارد، زیرا میانه ی 1 ها بالای میانه ی 0 ها قرار میگیرد.



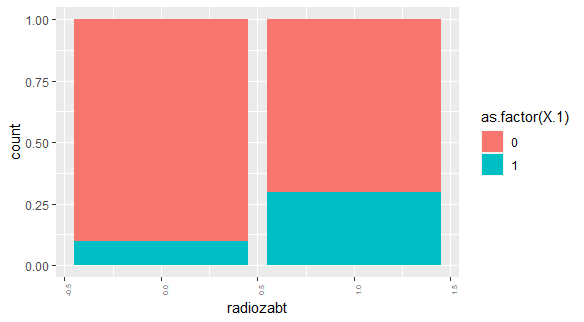
در پیش گوی ماشین داشتن، اگر 1 باشد با احتمال بیشتری برآمد 1 میشود و پیش گوی مناسبی برای تعیین برآمد است.(1263 نفر ماشین دارند و 1393 نفر ماشین ندارند)

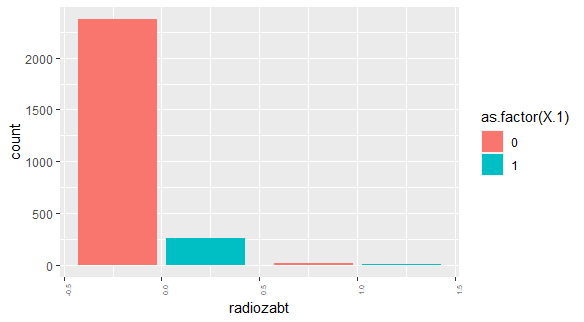


با 0 بودن این پیشگو(موتور نداشتن) احتمال بیشتری برای 1 بودن برآمد است. (740 نفر موتور دارند و 1916 نفر ندارند)

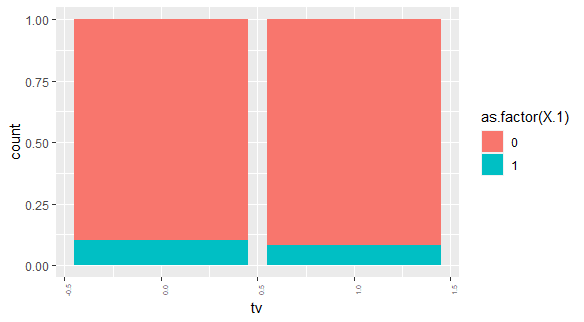
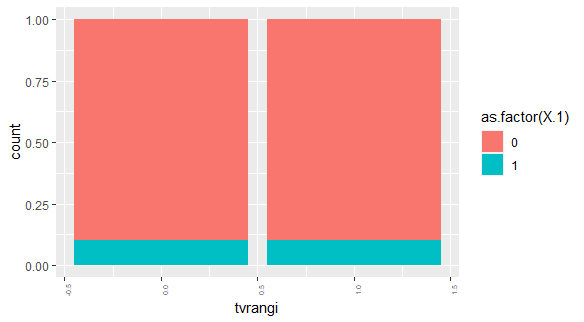


پیش گوی دوچرخه داشتن، پیش گوی مناسبی نیست زیرا هیچ تمیزی بین 0 و 1 ها ایجاد نمیکند.

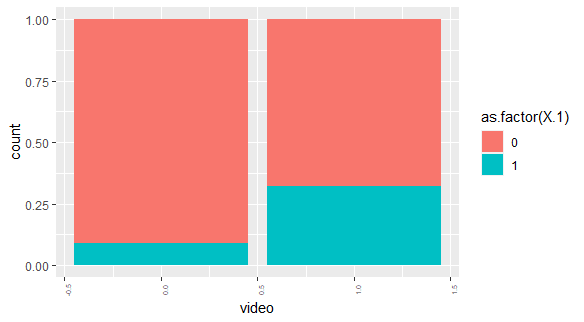




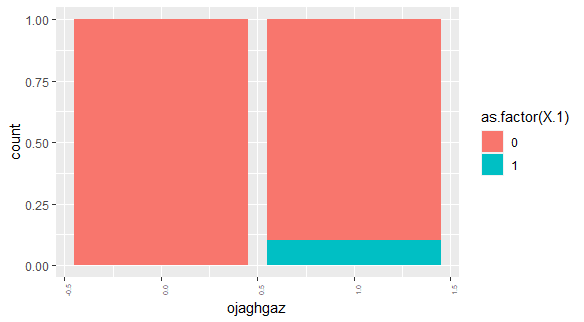
باتوجه به نمودار فوق رادیو ضبط به خوبی نمیتواند 0 و 1 های برآمد را تمیز دهد زیر تعداد آنهایی که متغییر پیش گوی آنها 1 است بسیار کم است درنتجیه کاندید حذف میباشد.



با توجه به نمودار های تلویزیون و تلویزیون رنگی، پیش گوی خوبی برای تمیز دادن برآمد نمیباشند و کاندید حذف هستند.

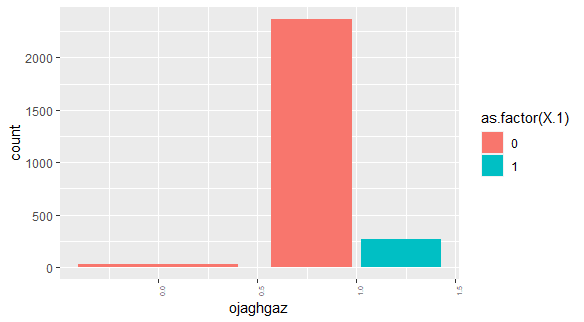


باتوجه به نمودار این پیش گو اگر 1 باشد تاثیر زیادی در 1شدن متغییر برآمد دارد.

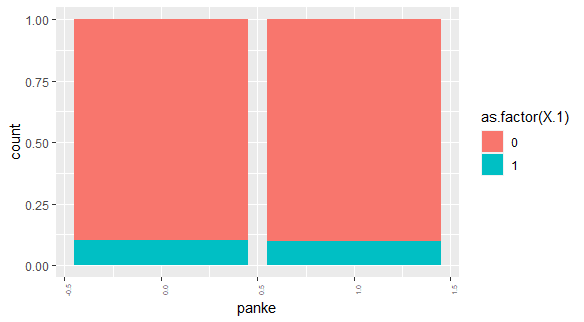


پیش گوی فوق به ظاهر برای تشخیص دادن 0 های برآمد مناسب است یعنی اگر خانواری اجاق گاز نداشت متغییر برآمد آن در رده ی 0 قرار میگیرد.

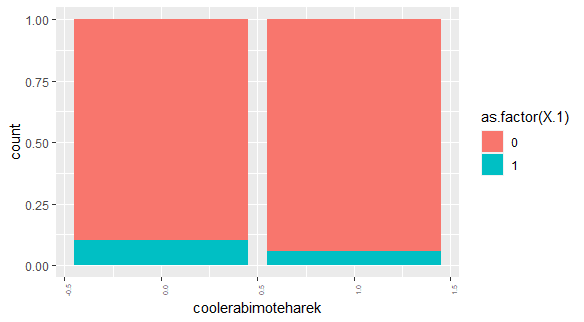
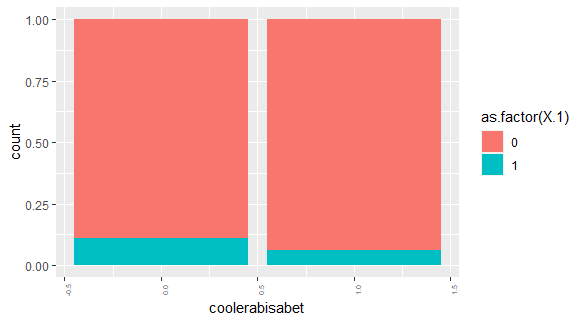
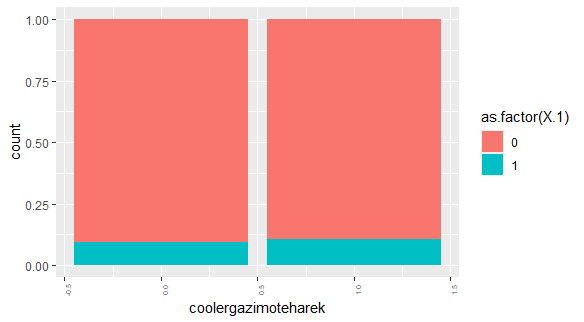
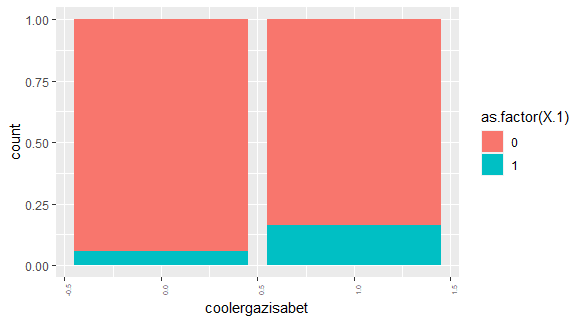
اما پس از بررسی مجدد و رسم نمودار زیر:



مشاهده شد که تعداد اشخاصی که اجاق گاز ندارند بسیار کم است(تنها 25 نفر) و درنتجیه پیش گوی مناسبی نمیباشد و کاندید حذف است.

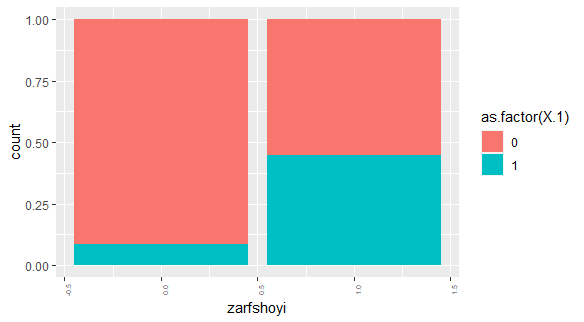


باتوجه به نمودار فوق، درمیابیم که داشتن یا نداشتن پنکه تاثیری در متغییر برآمد ندارد و کاندید حذف میباشد.

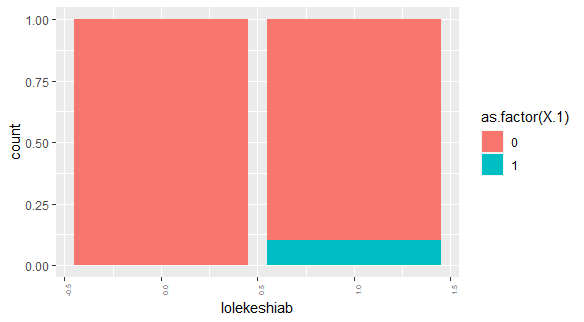


باتوجه به نمودار های فوق، کولر گازی ثابت بر روی متغییر برآمد تاثیر دارد و متغییر کولر گازی متحرک و کولر آبی و کولر آبی متحرک پیش گوی خوبی برای متغییر برآمد نیست و باید بیشتر مورد بررسی قرار بگیرند .

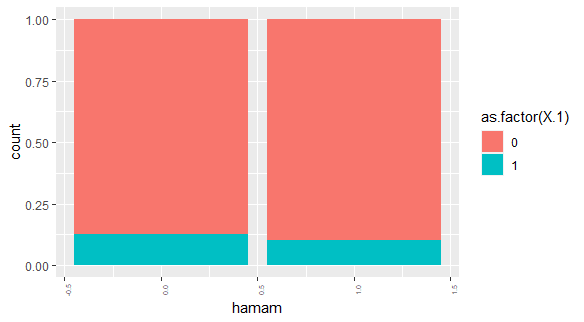
تعداد 1ها در کولرآبی متحرک 87 و تعداد 0 ها 2569 که به این دلیل پیش گوی خوبی نمیتواند باشد.



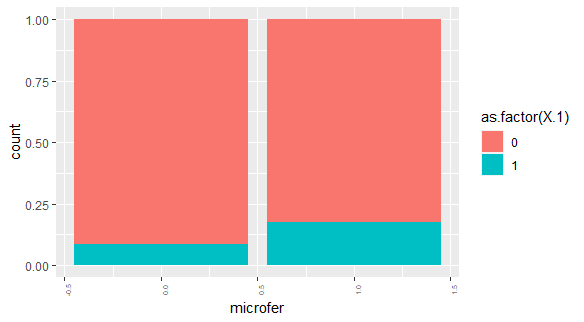
باتوجه به نمودار فوق، ماشین ظرفشویی پیش گوی خوبی برای تشخیص 1 ها متغییر برآمد است.



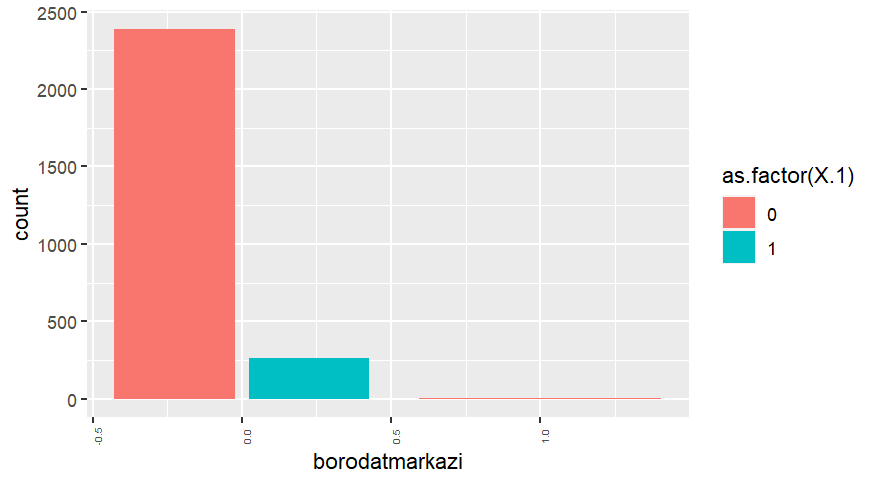
پیش گوی لوله کشی آب به ظاهر برای تشخیص 0 های متغییر برآمد مناسب است. ولی تنها 7 نفر از داده های ما لوله کشی آب ندارند پس این پیش گو کاندید حذف است.



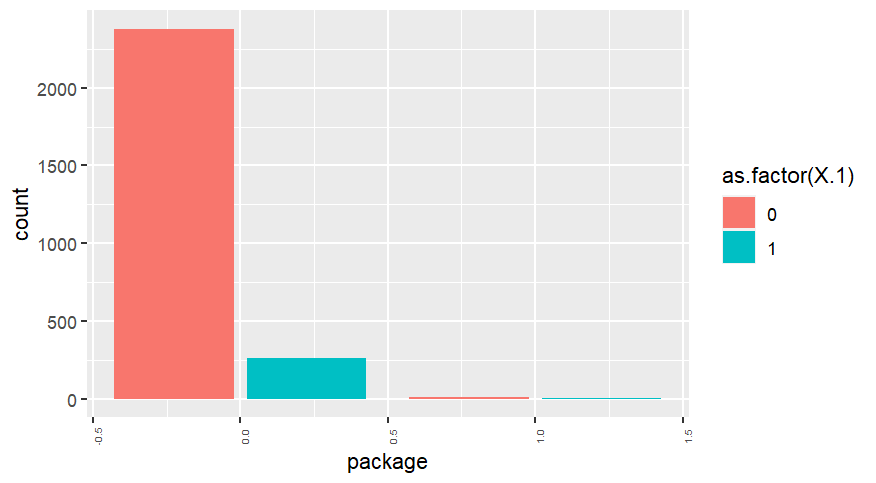
باتوجه به نمودار فوق، پیش گوی حمام نمیتواند به خوبی متغییر برآمد را تمیز دهد و درنتیجه کاندید حذف میباشد.



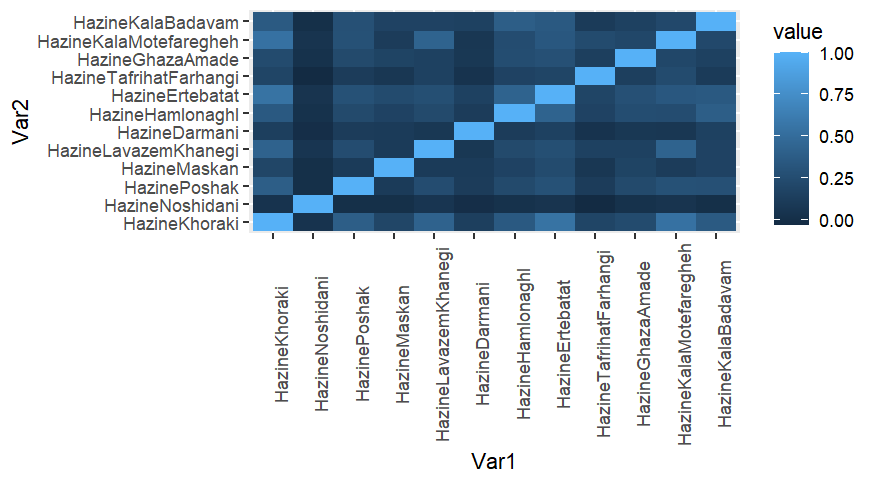
باتوجه به نمودار فوق، پیش گوی مایکروفر میتواند پیش گوی خوبی برای تمیز دادن متغییر برآمد باشد.



باتوجه به نمودار برودت مرکزی برحسب متغییر برآمد مشاهده میشود که تعداد 1 های این پیش گو بسیار کم است(2 عدد ثبت شده) و درنتیجه این پیش گو کاندید حذف میباشد.



باتوجه به نمودار متغییر برآمد برحسب پکیج، چون تنها آنهایی که پکیج دارند تعدادشان 14 نفر است نمیتواند پیش گوی خوبی باشد.



باتوجه به نمودار حرارتی برای پیش گو های هزینه، همبستگی زیادی بین متغییر های پیش گوی هزینه دیده نمیشود.

در نتیجه نمیتوان طبق این نمودار آنها را ادغام یا حذف کرد.

خلاصه ی نتیجه ی تصویری سازی و کاهش بعد

متغییر نوع خانوار طبق نمودار کاندید حذف بود حال با کمک گرفتن از خلاصه رگرسیون مشاهده شد این متغییر پیش گوی خوبی نمیتواند باشد زیرا

p-value آن 0.99 شد.

متغییر دو چرخه طبق نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.45 شد و درنتیجه در رگرسیون نیز معنا دار نیست.

متغییر رادیو ضبط در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.98 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

متغییر تلویزیون در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.88 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

متغییر تلویزیون رنگی در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.28 شد درنتیجه این متغییر را فعلا نگه میداریم.

متغییراجاق گاز در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.97 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

متغییر پنکه در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.22 شد و درنتیجه این متغییر را فعلا نگه میداریم.

متغییر کولر آبی متحرک در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز p-value آن 0.80 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

متغییرلوله کشی آب در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.98 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

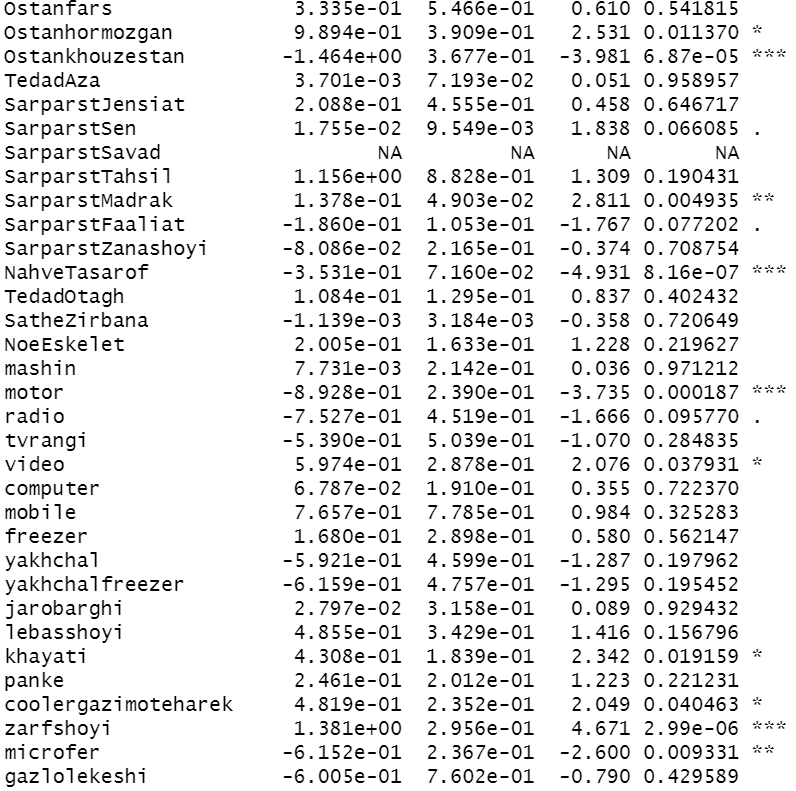
متغییر حمام در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

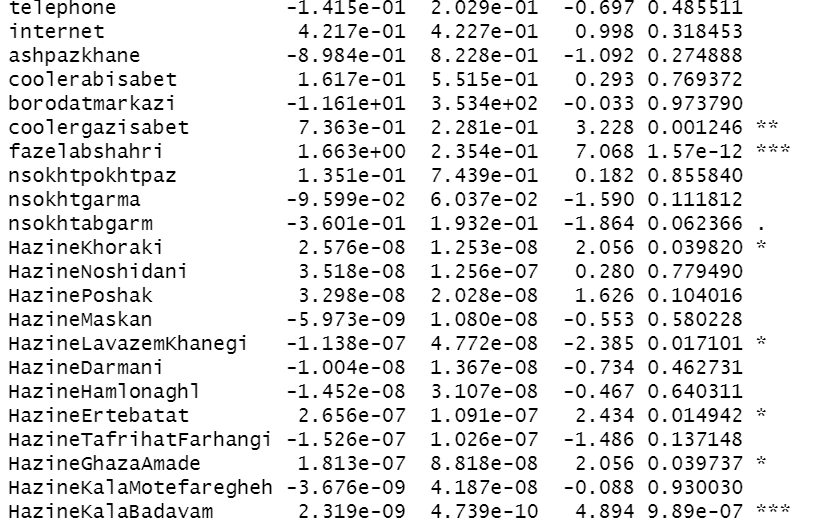
p-value آن 0.34 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

متغییرپکیج در نمودار ها کاندید حذف بود در خلاصه رگرسیون نیز

p-value آن 0.57 شد و درنتیجه پیش گوی خوبی نمیتواند باشد.

حال با توجه به تصویر زیر متغییر هایی که p-value آنها بالا 0.75 است را کاندید حذف قرار میدهیم.



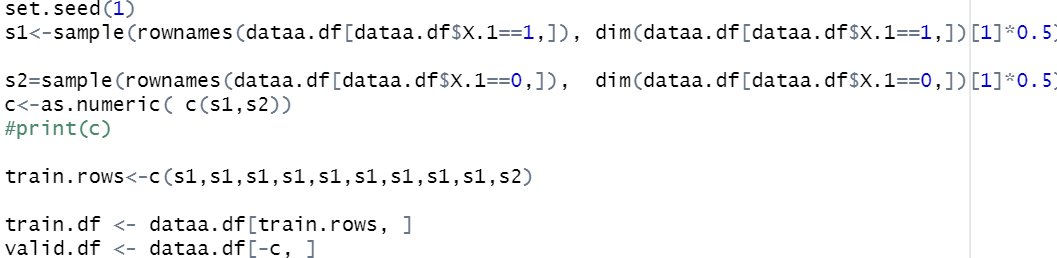


با توجه به تصویر فوق، متغییر های تعداد اعضا، ماشین، جارو برقی، کولرآبی ثابت(در نمودار ها نیز کاندید حذف بود)، برودت مرکزی(در نمودار ها نیز کاندید حذف بود)، نوع سوخت پخت پز، هزینه نوشیدنی و هزینه کالا متفرقه کاندید حذف میباشند.

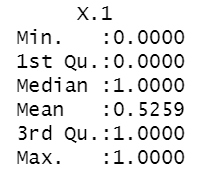
با حذف متغییر های فوق 45 متغییر پیش گو باقی میماند.

بیش نمونه گیری: (مهم ترین قسمت برای بالا بردن دقت)

میدانیم در متغییر برآمد تعداد 0 ها نه برابر 1 ها است برای اینکه این مشکل را برطرف کنیم از کد زیر استفاده میکنیم:

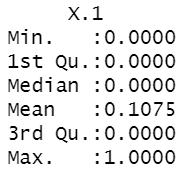


به اینصورت که پنجاه درصد از متغییر های برآمدی که 1 هستند را برداشته و پنجاه درصد از متغییر های برآمد را که 0 هستند بر میداریم سپس متغییر هایی که 1 هستند را 9 بار تکرار میکنیم(وزن 9 میدهیم) و در نتیجه فضای نمونه ی آموزشی ما متعادل میشود.



همانگونه که مشاهده میشود نسبت 0 ها و 1 ها تقریبا 50 به 50 شده است(میانگین 0.52).

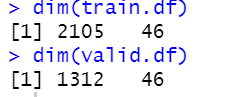
مجموعه ی اعتبار سنجی را نیز بدون وزن دهی و به همان نسبت مجموعه ی داده های اصلی انتخاب میکنیم.)میانگین 0.10)



در نتیجه با این کار مشکل نامتعادل بودن 0 و 1 ها برای مجموعه ی آموزشی را حل نموده ایم.

نکته ی بسیار مهم: باید دقت شود در هنگام محاسبه ی دقت بر روی مجموعه ی آموزشی مقدار دقت را اصلاح کنیم و به همون شکل داده های اصلی دربیاوریم، یعنی مقدار تشخیص دقت 0 ها(حساسیت) را با دادن وزن 9 محاسبه کنیم.

دقت شود که در افراز مجموعه ی آموزشی و اعتبار سنجی، 55 درصد برای مجموعه ی آموزشی(که 1 های آن را نه برابر کردیم) و 45 درصد از داده های برای مجموعه آموزشی انتخاب شده است.

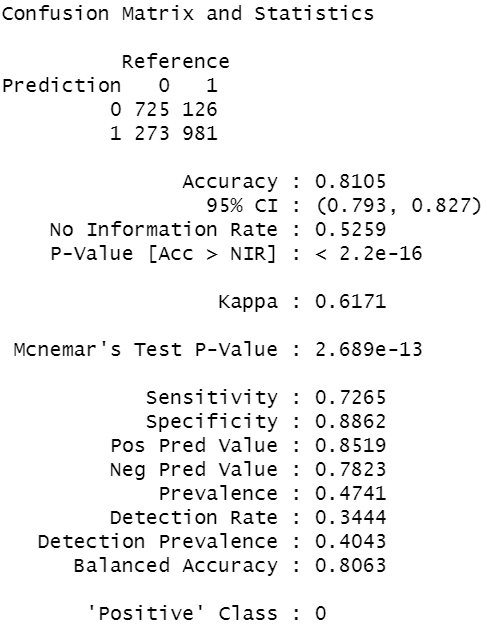


مدل رگرسیون لجستیک:

مدل رگرسیون را بر روی داده های بیش نمونه گیری شده اجرا میکنیم این مدل با مقدار برش 0.5 دقت 0.80 میدهد ولی دقت مشخص ساز 0.82 است و دقت حساسیت 0.78 است.

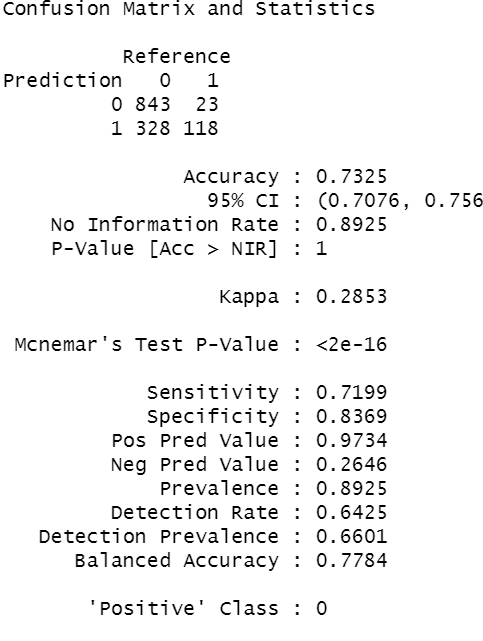
با مقدار برش 0.4 دقت 0.75 میشود ولی دقت مشخص ساز 0.88 و دقت حساسیت آن 0.73 میشود و چون شناسایی 1 ها برای ما ارزش بیشتری دارد این مقدار برش را انتخاب میکنیم.

دقت مدل بر روی داده های آموزشی برابر است با:



حال باید این دقت را با وزن 9 دادن (با توجه به توضیحات قسمت بیش نمونه گیری) به حساسیت اصلاح کرد که در این صورت دقت بر روی داده های آموزشی برابر میشود با 0.742.

حال ماتریس درهم ریختگی مدل بر روی داده های اعتبار سنجی را در شکل زیر میبینیم:

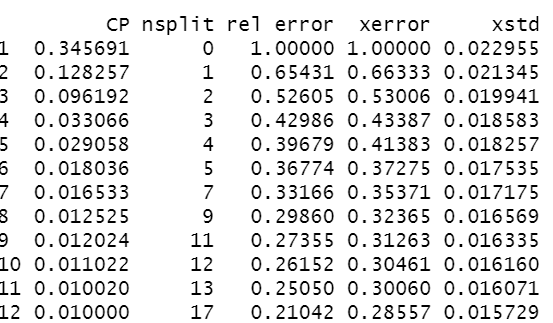


که دقت بر روی داده های اعتبار سنجی برابر است با 0.732 و چون تفاوت معنا داری با دقت بر روی داده های آموزشی ندارد از خطر بیش برازش در امانیم.

AIC این مدل برابر است با 1814.4 که عدد بالایی است.

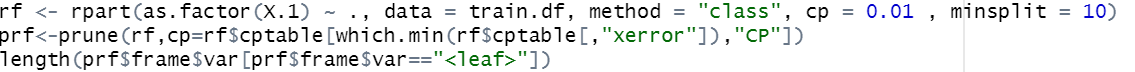
مدل درخت تصمیم:

مدل درخت را بر روی داده های آموزشی خود برازش میکنیم و سپس cp را برای هرس درخت انتخاب میکنیم که کمترین خطا را بر روی داده ها اعتبار سنجی داشته باشد باتوجه به شکل زیر داریم:

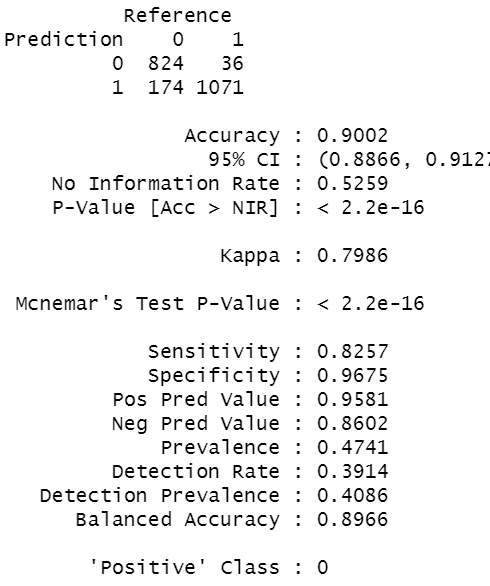


سطر دوازدهم کمترین خطا بر روی داده های اعتبار سنجی مقاطع را دارد.

با دستور زیر درخت با بهترین هرس را انتخاب میکنیم (زیرا درختمان بدون هرس به شدت دچار بیش برازش بود):

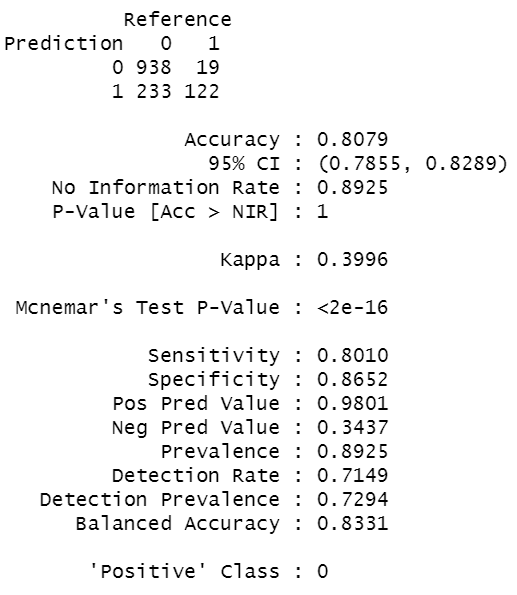


حال ماتریس درهم ریختگی بر روی مجموعه ی آموزشی محاسبه میکنیم که برابر است با:



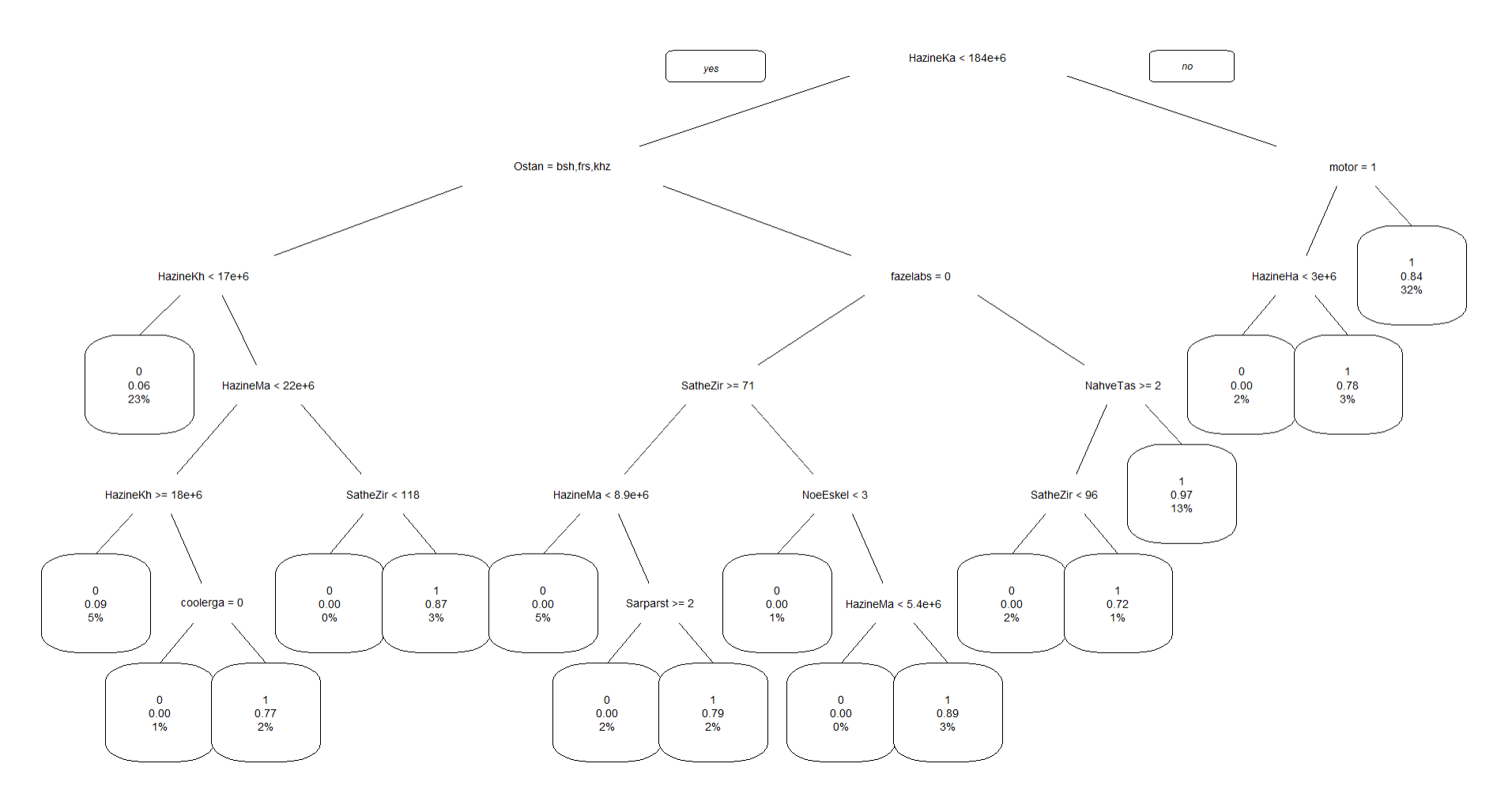
با توجه به ماتریس درهم ریختگی فوق با اصلاح دقت با دادن وزن 9 به حساسیت، دقت اصلاح شده برابر است با 0.839.

حال باتوجه به ماتریس درهم ریختگی بر روی داده های اعتبار سنجی داریم:

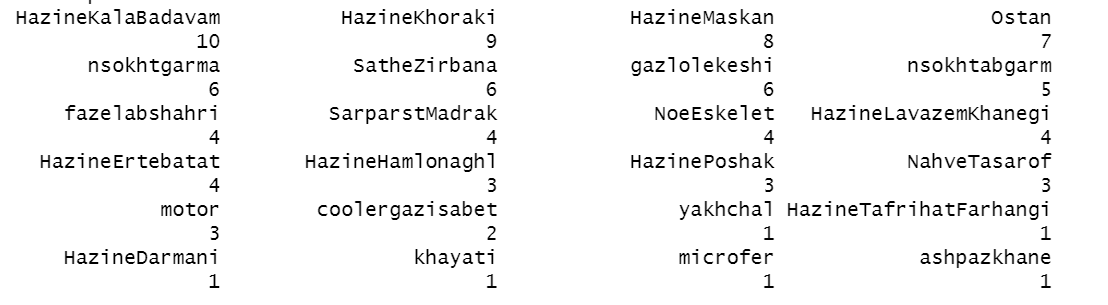


که دقت بر روی داده های اعتبار سنجی برابر است با 0.808 و دقت مشخص سازی برابر است با 0.865 که به خوبی 1 ها تمیز داده میشوند و چون دقت مدل بر روی داده های اعتبار سنجی تفاوت معنا داری با دقت بر روی داده های آموزشی ندارد از خطر بیش برازش در امانیم.

درخت هرس شده در شکل زیر نشان داده شده است.

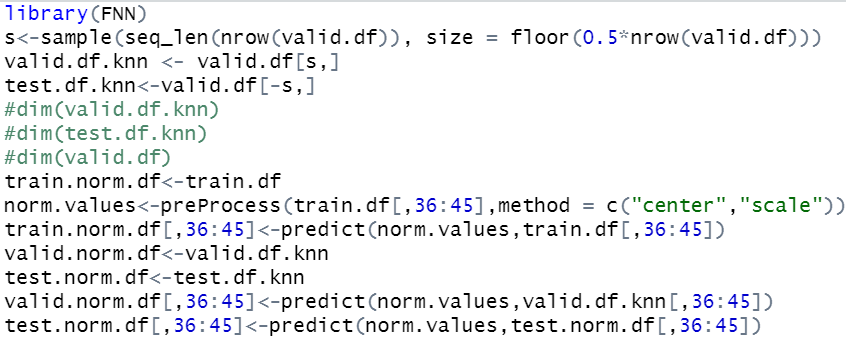


و متغییر هایی که در ساخت درخت موثر بوده اند در شکل زیر نشان داده شده اند:

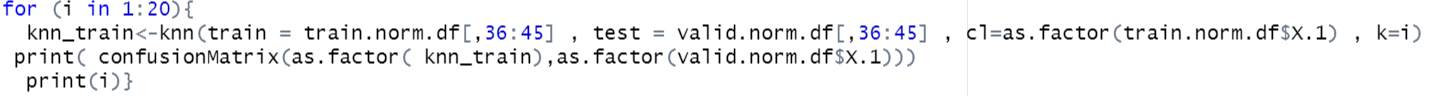


مدل k-نزدیکترین همسایه:

میدانیم برای مدل نزدیک ترین همسایه باید متغییر های پیوسته را نرمال کنیم و همچنین برای انتخاب بهترین k نیاز به استفاده ی مجموعه ی اعتبار سنجی داریم درنتیجه مجموعه ی جدید تست را نیز برای محاسبه ی دقت مدل بسازیم درنتیجه با کد زیر هم مجموعه ی جدید تست را میسازیم و هم نرمال سازی را انجام میدهیم:



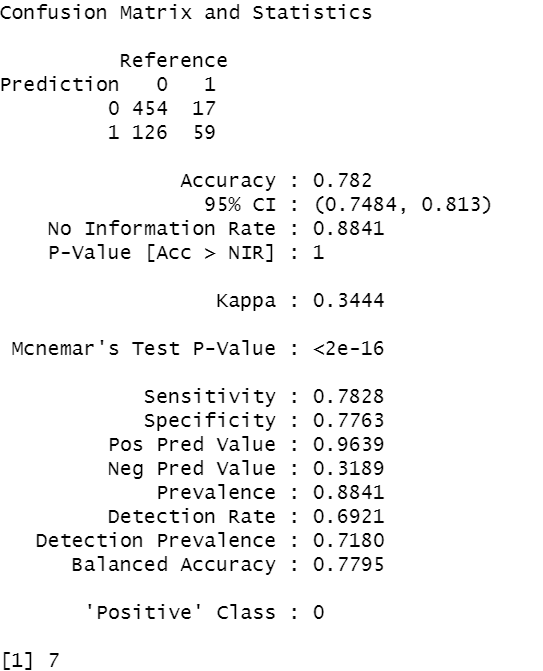
حال با استفاده از یک حلقه بهترین k را برای مدلمان طبق کد زیر انتخاب مینماییم:



یک نکته ی جالب که در داده ها اتفاق افتاد این است که با افزایش k حساسیت کاهش میابد و مشخص سازی افزایش میابد و چون وابستگی دقت مدل به جساسیت بیشتر است(نه برابر مشخص ساز تاثیر دارد) درنتیجه دقت مدل نیز کاهش میابد.

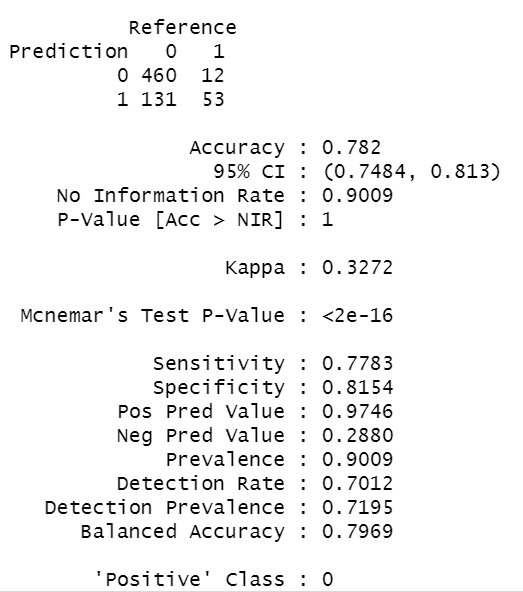
با استفاده از کد فوق بهترین k ای که ما برای مدل انتخاب میکنیم تا هم مشخص سازی حفظ شود هم دقت مدل 7 است.

ماتریس درهم ریختگی مدل بر روی داده های اعتبار سنجی برابر است با:



که دقت آن برابر است با 0.782 و دقت مشخص سازی برابر است با 0.776.

حال ماتریس درهم ریختگی بر روی داده های تست برابر است با:



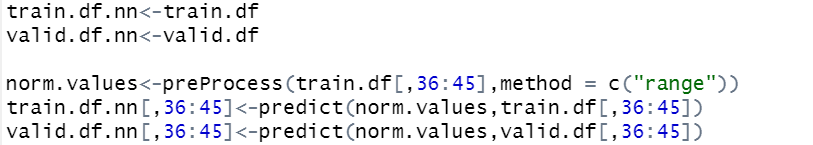
دقت بر روی این مدل برابر است با 0.782 و دقت مشخص سازی برابر است با 0.815.

مدل شبکه های عصبی:

برای استفاده از مدل شبکه های عصبی بهتر است ابتدا متغییر های پیوسته را به بازه ی [0,1] ببریم زیرا در غیر اینصورت شبکه ی عصبی با سه لایه ی پنهان دقتی برابر با 0.34 دارد.

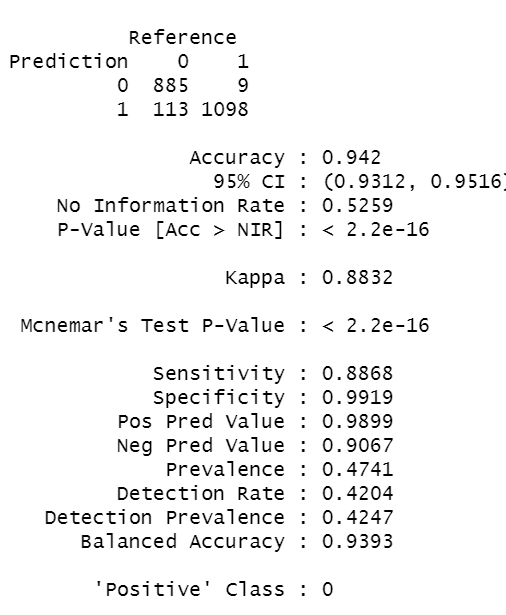
همچنین در این مدل از متغییر های غیر عددی نمیتوانیم بعنوان پیشگو استفاده کنیم.

با دستور زیر متغییر های پیوسته را به بازه [0,1] میبریم:



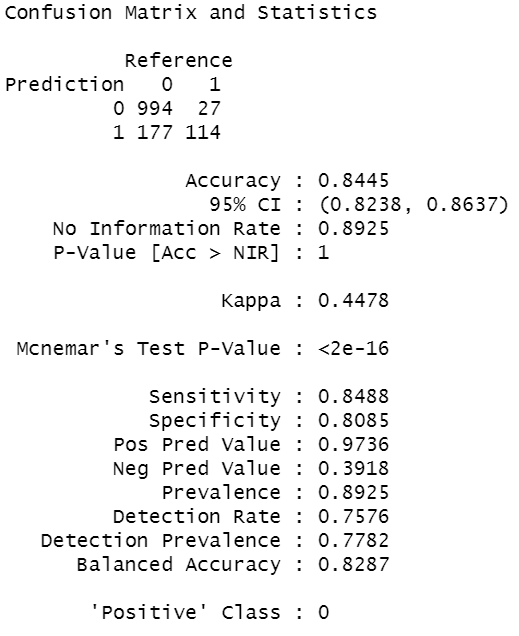
حال مدل را بر روی داده های نرمال شده با چهار لایه ی پنهان برازش مینماییم.

ماتریس درهم ریختگی برای مجموعه داده های آموزشی برابراست با:



با توجه به نکته گفته شده در قسمت بیش نمونه گیری باید وزن 9 را به حساسیت دهیم و مجدد دقت را محاسبه کنیم که در اینصورت دقت این مدل برابرست با 0.894.

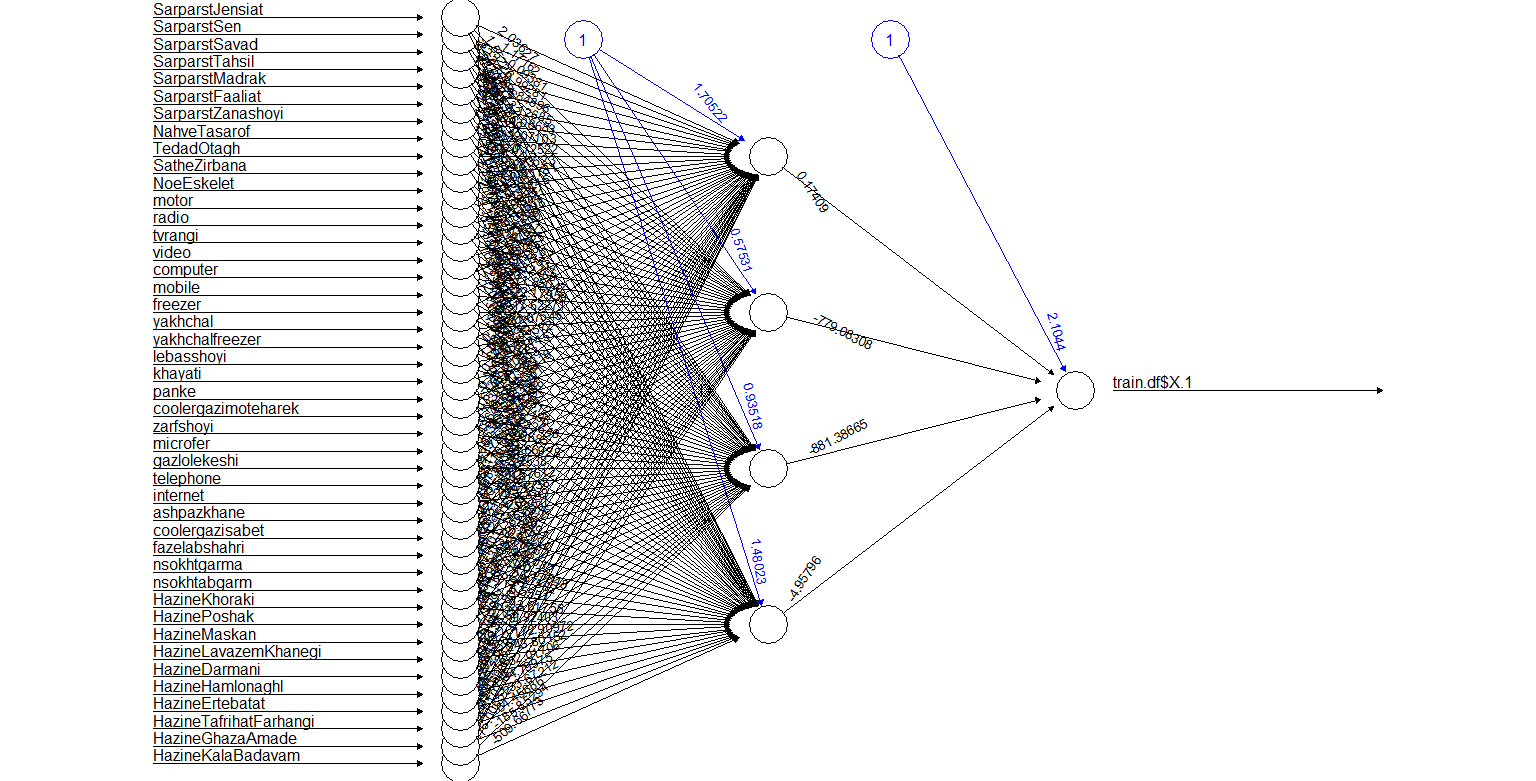
حال ماتریس درهم ریختگی بر روی داده های اعتبار سنجی برابرست با:



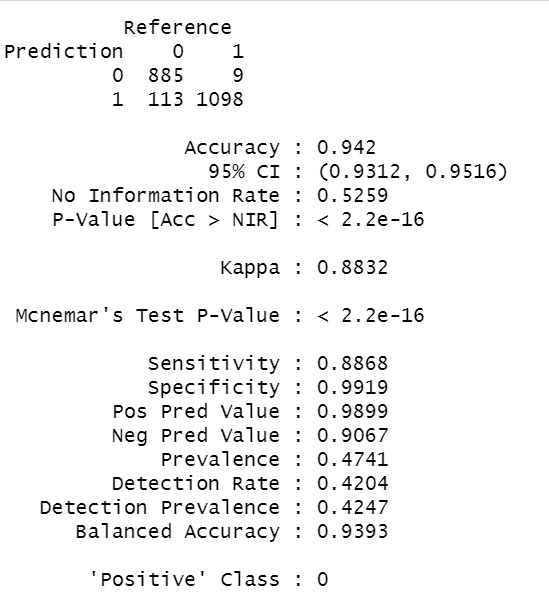
دقت این مدل بر روی داده های اعتبار سنجی برابر است با 0.844.

چون مجموعه ی اعتبار سنجی ما 45 درصد از کل داده هایمان را شامل میشود میتوان گفت از بیش برازش همچنان در امانیم.

شبکه ی عصبی با چهار لایه ی پنهان در شکل زیر آورده شده است.

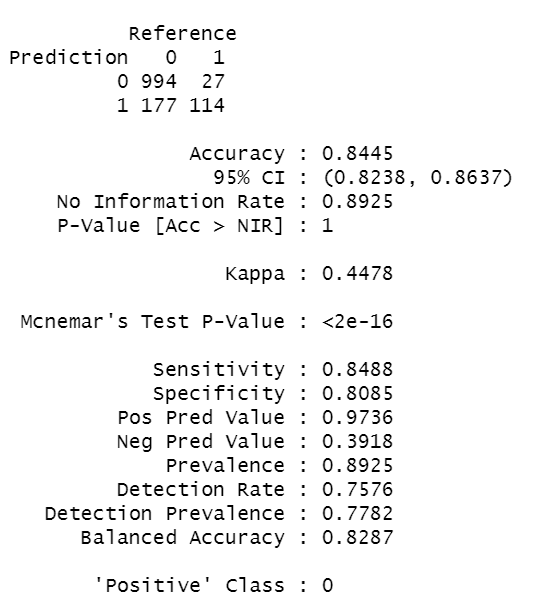
حال شبکه عصبی با پنج لایه پنهان را بر روی داده های آموزشیمان برازش میکنیم.

ماتریس درهم ریختگی شبکه عصبی با پنج لایه ی پنهان بر روی داده های آموزشی برابرست با:



باتوجه به نکته ی بیش نمونه گیری با وزن 9 داده به دقت حساسیت، دقت برابرست با 0.894.

ماتریس درهم ریختگی شبکه عصبی با پنج لایه ی پنهان بر روی داده های اعتبار سنجی برابرست با:



که دقت این مدل بر روی داده های اعتبار سنجی برابر است با 0.844.

پس از مشاهده شکل آن متوجه شدم یک لایه عصبی اضافه شده بیهوده است و دقیقا مشابه همان مدل با چهار لایه ی عصبی کار میکند.

انتخاب بهترین مدل:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| نام مدل | نزدیک ترین همسایه | درخت تصمیم | رگرسیون | شبکه های عصبی |
| دقت مدل | 0.782 | 0.808 | 0.732 | 0.844 |
| مشخص سازی | 0.815 | 0.865 | 0.836 | 0.808 |
| حساسیت | 0.778 | 0.801 | 0.719 | 0.848 |
| ایرادات |  | خطر بیش برازش | AIC بالا | خطر بیش برازش |

باتوجه به دقت مدل ها، مدل شبکه عصبی مدل مناسبی برای پیش گویی این مجموعه داده است. البته باید دقت شود که دچار بیش برازش در این مدل نشویم که در قسمت مدل شبکه عصبی در این باره توضیح داده شد.

باتشکر از توجه شما