

## پروژه داده کاوی

استاد راهنمای: دکتر محمدرضا فقیهی حبیب آبادی

گردآورنده: آرش سجادی، شماره دانشجویی: ۴۰۰۴۲۲۰۹۶

دانشکده ریاضی دانشگاه شهید بهشتی

۱۴۰۱ تیر

### چکیده:

داده‌های من در این پروژه مربوط به طرح آمارگیری هزینه و درامد خانوارهای شهری در سال ۱۳۹۹ است. داده‌های من توسط مرکز آمار ایران جمع آوری شده است. این داده شامل ۶۵ پیشگو<sup>۱</sup> چون خصوصیات اجتماعی اعضای خانوار، مشخصات محل سکونت، عناوین مختلف هزینه است. من در این پروژه ۲۳۳۱ خانوار را در سه استان تهران، البرز و قزوین مورد بررسی قرار خواهم داد. بنا به تعیین تکلیف کمیسیون تلفیق مجلس شورای اسلامی، در خصوص تبصره ۱۴ لایحه بودجه، به دهک<sup>۲</sup> دهم خانوارها یارانه‌ای تعلق نمی‌گیرد. از این رو هدف من در این پروژه رده‌بندی، با رده توفیق<sup>۳</sup> دهک دهم به منظور شناسایی خانوارهای این دهک است.

منبع اصلی من در کل این پروژه کتاب داده کاوی برای تحلیل خودکار کسب و کار: مفاهیم، فنون و کاربردهای R [۱] است. به علاوه منع من برای تولید واژه‌نامه، پس از واژه‌نامه منبع اصلی مذکور، واژه‌نامه رسمی انجمن ریاضی ایران [۲] و انجمن آمار ایران [۳] است.

**کلمات کلیدی:** داده کاوی، هزینه و درامد خانوارهای شهری

### فهرست مطالب

		۱ مقدمات و معرفی داده‌ها	
۱۰	تصویری سازی و اکتشاف داده‌ها	۲	
۱۰	بافت‌نگار ویژگی‌ها	۱-۲	
۱۴	نمودار چگالی	۲-۲	۱-۱ مقدمه
۱۵	نمودار میله‌ای	۳-۲	
۱۶	نمودار میله‌ای متغیرهای دودویی	۱-۳-۲	
۲۱	نمودار جعبه‌ای	۴-۲	۱-۱ هدف داده کاوی
۲۳	نمودار حرارتی	۵-۲	۲-۱ تعریف هر متغیر
۲۵	نمودار پراکنش سه بُعدی	۶-۲	۳-۱ خلاصه وضعیت ویژگی‌ها
۲۵	کاهش بُعد	۷-۲	۴-۱ پیش‌پردازش داده‌ها
۲۷	رگرسیون لجستیک	۳	
۳۳	مدل درخت تصمیم	۴	
۳۶	مدل k نزدیک‌ترین همسایه	۵	
۳۷	مدل شبکه عصبی	۶	
			۱-۴-۱ پاکسازی داده‌ها
			۲-۴-۱ مقادیر گم شده
			۳-۴-۱ افزای داده‌ها
			۴-۴-۱ دسته‌بندی ویژگی‌ها
			۵-۴-۱ نرمالیه کردن متغیرها
			۶-۴-۱ کدگذاری متغیرهای رسته‌ای

## انتخاب مدل نهایی

## مراجع

واژه نامه انگلیسی به فارسی

واژه نامه فارسی به انگلیسی

## مقدمه

داده<sup>۴</sup>ها من ثبت مشخصات ۲۳۳۱ خانوار شهری در استان های تهران، البرز و قزوین است. این داده ها توسط مرکز آمار ایران [۴] در طرح آمارگیری هزینه و درامد خانوارهای شهری و روستایی در سال ۱۳۹۹ جمع آوری شده است. هر ثبت<sup>۵</sup> شامل ۶۵ پیشگو است. ۲۶۵ ثبت از داده ها شامل مقادیر گم شده اند. در نتیجه ۲۰۶۶ داده سالم و بدون ابهام در اختیارم خواهد بود. البته تمام تلاش را می کنیم که از همه داده ها تا حد امکان استفاده کنیم.

## ۱ مقدمات و معرفی داده ها

هر ثبت در این پروژه مشخص کننده ویژگی های یک خانوار در یکی از شهرهای تهران، کرج و یا قزوین است که پرسشنامه مرکز آمار ایران را در طرح هزینه و درامد خانوار را در سال ۱۳۹۹ پر کرده است. در ادامه با جزئیات بیشتری به معرفی داده ها خواهیم پرداخت

### ۱-۱ هدف داده کاوی

با توجه به قانون تصمیمات جدیدی که در کشور در خصوص یارانه ها اجرا یافته است، یارانه به دهک دهم جامعه تعلق نخواهد گرفت. من نیز قصد دارم برآمدی رسته ای<sup>۶</sup> تعریف کنم که بیانگر تعلق گرفتن یا نگرفتن یارانه نقدی یا غیر نقدی به خانوارها است. به عبارت دیگر و یا حتی به زبان عامیانه هدف من در این پروژه تشخیص بودن یا نبودن خانوار در دهک برتر اقتصادی است.

### ۱-۲ تعریف هر متغیر

۱. نام استان: متغیر اول، یک متغیر رسته ای بیان کننده نام استانی است که داده در آنجا ثبت شده است. این سه رسته شامل استان های تهران، البرز و قزوین هستند.

۲. نوع خانوار: این ویژگی<sup>۷</sup>، یک ویژگی دودویی<sup>۸</sup> است که توصیف کننده نوع سکونت خانوار است. رده ۱ نمایانگر "عمولی ساکن" و رده ۲ نمایانگر "گروهی" است. (این

<sup>4</sup>Data

<sup>5</sup>Record

<sup>6</sup>Categorical

<sup>7</sup>Feature

<sup>8</sup>Binary

رسته ها برای تفکیک خانوارهایی که به صورت جمعی ساکن هستند و خانوارهایی که محل سکونت مستقل دارند در نظر گرفته شده است.)

### ۳. تعداد اعضای خانوار

۴. جنسیت سرپرست: پیشگوی سرپرست جنسیت، نمایانگر جنسیت سرپرست خانوار است. رده ۱ برای خانوارهایی که سرپرست آقا دارد و رده ۲ برای خانوارهایی که سرپرست خانم دارد در نظر گرفته شده است.

### ۵. سن سرپرست

۶. سواد سرپرست خانوار: رده ۱ برای سرپرست خانوارهای با سواد و رده ۲ برای سرپرست خانوارهای بی سواد در نظر گرفته شده است.

۷. اشتغال به تحصیل سرپرست خانوار: آیا در حال حاضر تحصیل می کند؟ ۱-بلی ۲-خیر

۸. مدرک تحصیلی سرپرست خانوار: این پیشگو متغیر رسته ای است. دوره یا مدرک تحصیلی ابتدایی / سوادآموزی با رسته ۱، راهنمایی / متوسطه اول با رسته ۲، متوسطه / متوسطه دوم با رسته ۳، دیپلم و پیش دانشگاهی با رسته ۴، فوق دیپلم / کاردانی با رسته ۵، لیسانس / کارشناسی با رسته ۶، کارشناسی ارشد و دکترای حرفه ای با رسته ۷، دکترای تخصصی با رسته ۸، سایر و غیررسمی با رسته ۹ مشخص شده اند.

۹. فعالیت سرپرست: متغیر رسته ای به شرح: ۱-شاغل ۲- بیکار یا جویای کار ۳-دارای درامد بدون کار ۴-محصل ۵-خانه دار ۶-سایر

۱۰. وضعیت زناشویی سرپرست: متغیر رسته ای ۱-دارای همسر ۲-بی همسر بر اثر فوت همسر ۳-بی همسر بر اثر طلاق ۴- هرگز ازدواج نکرده

۱۱. نحوه تصرف محل زندگی؛ ۱-ملکی عرصه و اعیان؛ خانوار مالک زمین و بنای منزل سکونتی خود است. ۲-ملکی اعیان؛ خانوار تنها مالک بنای منزل سکونتی خود است. ۳-اجاری؛ خانوار منزل سکونتی خود را به ازای پرداخت مقداری پول به صورت قرض الحسن به مالک برای مدت معینی تصرف کرده است. ۴-در برابر خدمت؛ خانوار منزل سکونتی خود را در مقابل انجام کار یک یا چند نفر از اعضا یاش، تصرف کرده است. ۵-رایگان؛ هیچ یک از اعضا خانوار مبلغ یا خدمتی را برای منزل خود نمی پردازند و نه مالک زمین و نه بنای منزل سکونتی خود، هستند.

حذف خواهم کرد چراکه حاوی اطلاعات<sup>۹</sup> خاصی برای من نیست.)

۳۸. دسترسی به لوله کشی آب: (این متغیر رسته‌ای دودویی است اما از آنجا که در داده‌های من همه‌ی ثبت‌ها در این ویژگی رسته ۱ را ثبت کرده‌اند این ویژگی را از داده‌ها حذف خواهم کرد چراکه حاوی اطلاعات خاصی برای من نیست.)

۳۹. دسترسی به برق: (این متغیر رسته‌ای دودویی است اما از آنجا که در داده‌های من همه‌ی ثبت‌ها در این ویژگی رسته ۱ را ثبت کرده‌اند این ویژگی را از داده‌ها حذف خواهم کرد چراکه حاوی اطلاعات خاصی برای من نیست.)

۴۰. دسترسی به گاز لوله کشی

۴۱. دسترسی به تلفن

۴۲. دسترسی به اینترنت

۴۳. دسترسی به حمام شخصی

۴۴. داشتن آشپزخانه در منزل

۴۵. داشتن کولر آبی ثابت

۴۶. داشتن سیستم برودت مرکزی

۴۷. داشتن سیستم حرارت مرکزی

۴۸. داشتن پکیج

۴۹. داشتن کولرگازی ثابت

۵۰. داشتن فاضلاب شهری

۵۱. نوع سوخت مصرفی برای پخت و پز: برای این متغیر رسته‌ای ۱۰ رسته مختلف در پرسشنامه در نظر گرفته شده بود اما تمامی ثبت‌های من یا از گاز مایع (رسته ۳) و یا از گاز شهری (رسته ۴) استفاده می‌کنند. لذا می‌توانم همه درایه‌های این ویژگی را با صفر و یک نمایش دهم.

۵۲. نوع سوخت مصرفی برای گرمایش: برای این متغیر رسته‌ای ۱۰ رسته مختلف در پرسشنامه در نظر گرفته شده بود اما تمامی ثبت‌های من یا از گاز مایع (رسته ۱۳) و یا از گاز شهری (رسته ۱۴) استفاده می‌کنند. لذا می‌توانم همه درایه‌های این ویژگی را با صفر و یک نمایش دهم.

۱۲. تعداد اتاق

۱۳. مساحت زیربنا

۱۴. نوع اسکلت: متغیر رسته‌ای ۱-فلزی ۲-بتن‌آرمه ۳-سایر

۱۵. آیا از اتومبیل شخصی استفاده می‌کند؟

۱۶. آیا از موتورسیکلت شخصی استفاده می‌کند؟

۱۷. آیا از دوچرخه شخصی استفاده می‌کند؟

۱۸. آیا از رادیو استفاده می‌کند؟

۱۹. آیا از رادیو ضبط استفاده می‌کند؟

۲۰. آیا از تلویزیون استفاده می‌کند؟

۲۱. آیا از تلویزیون رنگی استفاده می‌کند؟

۲۲. آیا از ویدیو استفاده می‌کند؟

۲۳. آیا از کامپیوتر استفاده می‌کند؟

۲۴. آیا از موبایل استفاده می‌کند؟

۲۵. آیا از فریزر استفاده می‌کند؟

۲۶. آیا از یخچال استفاده می‌کند؟

۲۷. آیا از یخچال فریزر استفاده می‌کند؟

۲۸. آیا از اجاق گاز استفاده می‌کند؟

۲۹. آیا از جاروبرقی استفاده می‌کند؟

۳۰. آیا از ماشین لباسشویی استفاده می‌کند؟

۳۱. آیا از چرخ خیاطی استفاده می‌کند؟

۳۲. آیا از پنکه استفاده می‌کند؟

۳۳. آیا از کولر آبی متحرک استفاده می‌کند؟

۳۴. آیا از کولرگازی متحرک استفاده می‌کند؟

۳۵. آیا از ماشین ظرفشویی استفاده می‌کند؟

۳۶. آیا از مایکروویو و فرهای هالوژن دار استفاده می‌کند؟

۳۷. آیا این طور است که از هیچ‌یک از وسایل پرسیده شده در ویژگی‌های ۱۵ تا ۳۶ استفاده نمی‌کند؟ (این متغیر رسته‌ای دودویی است اما از آنجا که در داده‌های من همه‌ی ثبت‌ها در این ویژگی رسته ۰ را ثبت کرده‌اند این ویژگی را از داده‌ها

9		
10	TedadAza	SarparstJensiat
11	Min. : 1.000	Min. : 1.000
12	1st Qu.: 2.000	1st Qu.: 1.000
13	Median : 3.000	Median : 1.000
14	Mean : 3.224	Mean : 1.146
15	3rd Qu.: 4.000	3rd Qu.: 1.000
16	Max. : 14.000	Max. : 2.000
17		
18	SarparstSen	SarparstSavad
19	Min. : 20.00	Min. : 1.000
20	1st Qu.: 40.00	1st Qu.: 1.000
21	Median : 50.00	Median : 1.000
22	Mean : 51.96	Mean : 1.114
23	3rd Qu.: 62.00	3rd Qu.: 1.000
24	Max. : 97.00	Max. : 2.000
25		
26	SarparstTahsil	SarparstMadراك
27	Min. : 1.000	Min. : 1.000
28	1st Qu.: 2.000	1st Qu.: 1.000
29	Median : 2.000	Median : 4.000
30	Mean : 1.987	Mean : 3.339
31	3rd Qu.: 2.000	3rd Qu.: 4.000
32	Max. : 2.000	Max. : 9.000
33	NA's : 265	NA's : 265
34		
35	SarparstFaaliat	SarparstZanashoyi
36	Min. : 1.000	Min. : 1.00
37	1st Qu.: 1.000	1st Qu.: 1.00
38	Median : 1.000	Median : 1.00
39	Mean : 1.928	Mean : 1.25
40	3rd Qu.: 3.000	3rd Qu.: 1.00
41	Max. : 6.000	Max. : 4.00
42		
43	NahveTasarof	Tedad0tagh
44	Min. : 1.000	Min. : 1.000
45	1st Qu.: 1.000	1st Qu.: 3.000
46	Median : 1.000	Median : 4.000
47	Mean : 2.103	Mean : 3.641
48	3rd Qu.: 3.000	3rd Qu.: 4.000
49	Max. : 6.000	Max. : 9.000
50		
51	SatheZirbana	NoeEskelet
52	Min. : 9.00	Min. : 1.000
53	1st Qu.: 64.00	1st Qu.: 1.000
54	Median : 80.00	Median : 2.000
55	Mean : 85.94	Mean : 1.747
56	3rd Qu.: 100.00	3rd Qu.: 2.000
57	Max. : 360.00	Max. : 3.000
58		
59	mashin	motor
60	Min. : 0.0000	Min. : 0.00000
61	1st Qu.: 0.0000	1st Qu.: 0.00000

۵۳. نوع سوخت مصرفی برای تهیه: برای این متغیر رسته‌ای ۱۰ رسته مختلف در پرسشنامه در نظر گرفته شده بود اما تمامی ثبت‌های من یا از گاز مایع (رسته ۲۳) و یا از گاز شهری (رسته ۲۴) استفاده می‌کنند. لذا می‌توانم همه درایه‌های این ویژگی را با صفر و یک نمایش دهم.

۵۴. هزینه خوراک در ماه گذشته (برحسب ریال)

۵۵. هزینه نوشیدنی در ماه گذشته (برحسب ریال)

۵۶. هزینه پوشاش در ماه گذشته (برحسب ریال)

۵۷. هزینه مسکن در ماه گذشته (برحسب ریال)

۵۸. هزینه لوازم خانگی در ماه گذشته (برحسب ریال)

۵۹. هزینه درمان در ماه گذشته (برحسب ریال)

۶۰. هزینه حمل و نقل در ماه گذشته (برحسب ریال)

۶۱. سایر هزینه‌های مربوط به ارتباطات خانوار در ماه گذشته (برحسب ریال)

۶۲. هزینه مربوط به تفریحات فرهنگی در ماه گذشته (برحسب ریال)

۶۳. هزینه مربوط به خرید اغذیه آماده در ماه گذشته (برحسب ریال)

۶۴. هزینه مربوط به خرید کالای متفرقه در ماه گذشته (برحسب ریال)

۶۵. هزینه مربوط به خرید کالای بادوام در ماه گذشته (برحسب ریال)

### ۳-۱ خلاصه وضعیت ویژگی‌ها

در هنگام شروع کار با داده‌ها خوب است که خلاصه<sup>۱۰</sup> ای از شاخص‌های مرکزی مهم آنها بدانیم. من کلدها به همراه نتایج حاصل شده از آنها را که با مترجم<sup>۱۱</sup> زبان برنامه‌نویسی R اجرا شده است را در ادامه قرار خواهم داد.

```
1 > summary(X)
2   Data[, 2]           NoeKhanevar
3   Length:2331         Min. :1.000
4   Class :character    1st Qu.:1.000
5   Mode  :character    Median :1.000
6                               Mean  :1.002
7                               3rd Qu.:1.000
8                               Max. :2.000
```

<sup>10</sup>Summary

<sup>11</sup>Compiler

115	jarobarghi	lebasshoyi						
116	Min.	:0.0000	Min.	:0.0000				
117	1st Qu.	:1.0000	1st Qu.	:1.0000				
118	Median	:1.0000	Median	:1.0000				
119	Mean	:0.9605	Mean	:0.9185				
120	3rd Qu.	:1.0000	3rd Qu.	:1.0000				
121	Max.	:1.0000	Max.	:1.0000				
122								
123	khayati	panke						
124	Min.	:0.0000	Min.	:0.0000				
125	1st Qu.	:0.0000	1st Qu.	:0.0000				
126	Median	:0.0000	Median	:0.0000				
127	Mean	:0.4127	Mean	:0.0858				
128	3rd Qu.	:1.0000	3rd Qu.	:0.0000				
129	Max.	:1.0000	Max.	:1.0000				
130								
131	coolerabimoteharek	coolergazimoteharek						
132	Min.	:0.0000	Min.	:0.000000				
133	1st Qu.	:0.0000	1st Qu.	:0.000000				
134	Median	:0.0000	Median	:0.000000				
135	Mean	:0.0133	Mean	:0.003432				
136	3rd Qu.	:0.0000	3rd Qu.	:0.000000				
137	Max.	:1.0000	Max.	:1.000000				
138								
139	zarfshoyi	microfer						
140	Min.	:0.0000	Min.	:0.0000				
141	1st Qu.	:0.0000	1st Qu.	:0.0000				
142	Median	:0.0000	Median	:0.0000				
143	Mean	:0.1085	Mean	:0.1699				
144	3rd Qu.	:0.0000	3rd Qu.	:0.0000				
145	Max.	:1.0000	Max.	:1.0000				
146								
147	hichkodam	lolekeshiab						
148	Min.	:0	Min.	:1				
149	1st Qu.	:0	1st Qu.	:1				
150	Median	:0	Median	:1				
151	Mean	:0	Mean	:1				
152	3rd Qu.	:0	3rd Qu.	:1				
153	Max.	:0	Max.	:1				
154								
155	bargh	gazlolekeshi						
156	Min.	:1	Min.	:0.0000				
157	1st Qu.	:1	1st Qu.	:1.0000				
158	Median	:1	Median	:1.0000				
159	Mean	:1	Mean	:0.9996				
160	3rd Qu.	:1	3rd Qu.	:1.0000				
161	Max.	:1	Max.	:1.0000				
162								
163	telephone	internet						
164	Min.	:0.0000	Min.	:0.0000				
165	1st Qu.	:0.0000	1st Qu.	:1.0000				
166	Median	:1.0000	Median	:1.0000				
167	Mean	:0.7147	Mean	:0.7894				
62	Median	:1.0000	Median	:0.00000				
63	Mean	:0.5337	Mean	:0.09953				
64	3rd Qu.	:1.0000	3rd Qu.	:0.00000				
65	Max.	:1.0000	Max.	:1.00000				
66								
67	docharkhe	radio						
68	Min.	:0.0000	Min.	:0.000000				
69	1st Qu.	:0.0000	1st Qu.	:0.000000				
70	Median	:0.0000	Median	:0.000000				
71	Mean	:0.0961	Mean	:0.009438				
72	3rd Qu.	:0.0000	3rd Qu.	:0.000000				
73	Max.	:1.0000	Max.	:1.000000				
74								
75	radiozabt	tv						
76	Min.	:0.00000	Min.	:0.000000				
77	1st Qu.	:0.00000	1st Qu.	:0.000000				
78	Median	:0.00000	Median	:0.000000				
79	Mean	:0.07207	Mean	:0.002574				
80	3rd Qu.	:0.00000	3rd Qu.	:0.000000				
81	Max.	:1.00000	Max.	:1.000000				
82								
83	tvrangi	video						
84	Min.	:0.0000	Min.	:0.0000				
85	1st Qu.	:1.0000	1st Qu.	:0.0000				
86	Median	:1.0000	Median	:0.0000				
87	Mean	:0.9858	Mean	:0.3157				
88	3rd Qu.	:1.0000	3rd Qu.	:1.0000				
89	Max.	:1.0000	Max.	:1.0000				
90								
91	computer	mobile						
92	Min.	:0.0000	Min.	:0.0000				
93	1st Qu.	:0.0000	1st Qu.	:1.0000				
94	Median	:0.0000	Median	:1.0000				
95	Mean	:0.3308	Mean	:0.9674				
96	3rd Qu.	:1.0000	3rd Qu.	:1.0000				
97	Max.	:1.0000	Max.	:1.0000				
98								
99	freezer	yakhchal						
100	Min.	:0.0000	Min.	:0.0000				
101	1st Qu.	:0.0000	1st Qu.	:0.0000				
102	Median	:0.0000	Median	:0.0000				
103	Mean	:0.1802	Mean	:0.2437				
104	3rd Qu.	:0.0000	3rd Qu.	:0.0000				
105	Max.	:1.0000	Max.	:1.0000				
106								
107	yakhchalfreezer	ojaghgaz						
108	Min.	:0.0000	Min.	:0.000				
109	1st Qu.	:1.0000	1st Qu.	:1.000				
110	Median	:1.0000	Median	:1.000				
111	Mean	:0.7619	Mean	:0.991				
112	3rd Qu.	:1.0000	3rd Qu.	:1.000				
113	Max.	:1.0000	Max.	:1.000				
114								

221	1st Qu.:	0	1st Qu.:	0
222	Median :	0	Median :	0
223	Mean :	419233	Mean :	1453133
224	3rd Qu.:	0	3rd Qu.:	945000
225	Max. :	45000000	Max. :	75800000
226				
227	HazineMaskan		HazineLavazemKhanegi	
228	Min. :	1158000	Min. :	0
229	1st Qu.:	12499000	1st Qu.:	500000
230	Median :	20920000	Median :	845000
231	Mean :	28653389	Mean :	1206621
232	3rd Qu.:	33443000	3rd Qu.:	1332500
233	Max. :	801250000	Max. :	35600000
234				
235	HazineDarmani		HazineHamlonaghl	
236	Min. :	0	Min. :	0
237	1st Qu.:	0	1st Qu.:	745000
238	Median :	750000	Median :	1650000
239	Mean :	3176284	Mean :	2491068
240	3rd Qu.:	2600000	3rd Qu.:	3345000
241	Max. :	192000000	Max. :	80500000
242				
243	HazineErtebatat		HazineTafrihatFarhangi	
244	Min. :	0	Min. :	0
245	1st Qu.:	600000	1st Qu.:	0
246	Median :	1000000	Median :	0
247	Mean :	1182938	Mean :	183855
248	3rd Qu.:	1500000	3rd Qu.:	0
249	Max. :	10500000	Max. :	22800000
250				
251	HazineGhazaAmade		HazineKalaMotefaregheh	
252	Min. :	0	Min. :	0
253	1st Qu.:	0	1st Qu.:	450000
254	Median :	0	Median :	800000
255	Mean :	450149	Mean :	1196962
256	3rd Qu.:	0	3rd Qu.:	1302000
257	Max. :	30000000	Max. :	74360000

## ۴-۱ پیش‌پردازش داده‌ها

پیش‌پردازش<sup>۱۲</sup> داده به مراحلی گفته می‌شود که در آن داده‌ها برای داده‌کاوی آماده می‌شود. لازم به ذکر است که این مراحل جزء مهم‌ترین گام‌ها در داده‌کاوی هستند. در این بخش با استفاده از ماتریس<sup>۱۳</sup> داده‌ها را نهایی کنیم تا در بدوزد شروع فصل بعد، تصویری سازی<sup>۱۴</sup> روی ماتریس نهایی داده‌ها صورت بگیرد.

### ۱-۴-۱ پاکسازی داده‌ها

نکته اب که در بدوزد پاکسازی داده‌ها با آن به رو به رو شدم این بود که بسیاری از پیشگوهای من از نظر فلسفی دودویی بودند ( فقط

168	3rd Qu.:	1.0000	3rd Qu.:	1.0000
169	Max. :	1.0000	Max. :	1.0000
170				
171	hamam		ashpazkhane	
172	Min. :	0.0000	Min. :	0.0000
173	1st Qu.:	1.0000	1st Qu.:	1.0000
174	Median :	1.0000	Median :	1.0000
175	Mean :	0.9987	Mean :	0.9966
176	3rd Qu.:	1.0000	3rd Qu.:	1.0000
177	Max. :	1.0000	Max. :	1.0000
178				
179	coolerabisabet		borodatmarkazi	
180	Min. :	0.0000	Min. :	0.00000
181	1st Qu.:	1.0000	1st Qu.:	0.00000
182	Median :	1.0000	Median :	0.00000
183	Mean :	0.8962	Mean :	0.05749
184	3rd Qu.:	1.0000	3rd Qu.:	0.00000
185	Max. :	1.0000	Max. :	1.00000
186				
187	hararatmarkazi		package	
188	Min. :	0.0000	Min. :	0.0000
189	1st Qu.:	0.0000	1st Qu.:	0.0000
190	Median :	0.0000	Median :	0.0000
191	Mean :	0.1789	Mean :	0.2111
192	3rd Qu.:	0.0000	3rd Qu.:	0.0000
193	Max. :	1.0000	Max. :	1.0000
194				
195	coolergazisabet		fazelabshahri	
196	Min. :	0.00000	Min. :	0.0000
197	1st Qu.:	0.00000	1st Qu.:	0.0000
198	Median :	0.00000	Median :	0.0000
199	Mean :	0.04419	Mean :	0.4655
200	3rd Qu.:	0.00000	3rd Qu.:	1.0000
201	Max. :	1.00000	Max. :	1.0000
202				
203	nsokhtpokhtpaz		nsokhtarma	
204	Min. :	3	Min. :	13
205	1st Qu.:	4	1st Qu.:	14
206	Median :	4	Median :	14
207	Mean :	4	Mean :	14
208	3rd Qu.:	4	3rd Qu.:	14
209	Max. :	4	Max. :	14
210				
211	nsokhtabgarm		HazineKhoraki	
212	Min. :	23	Min. :	0
213	1st Qu.:	24	1st Qu.:	8448200
214	Median :	24	Median :	11768600
215	Mean :	24	Mean :	13842026
216	3rd Qu.:	24	3rd Qu.:	16718100
217	Max. :	24	Max. :	203927000
218				
219	HazineNoshidani		HazinePoshak	
220	Min. :	0	Min. :	0

<sup>12</sup>Preprocessing

<sup>13</sup>Matrix

<sup>14</sup>Visualization

## ۲-۴-۱ مقادیر گم شده

<sup>۱۴</sup> در حال حاضر ۲۶۵ متغیر تهی<sup>۱۶</sup> یا به تعبیری مقدار گم شده<sup>۱۷</sup> در داده های من وجود دارد. من خیلی سریع متوجه شدم هیچ ایرادی در این داده ها نیست و در واقع از نظر مفهومی این ها متغیر تهی نیستند. داستان به این شکل است که یک بار با سواد بودن و یا نبودن سرپرست خانوارها پرسیده شده. برای افراد با سواد، دو سؤال دیگر نیز در پرسشنامه وجود داشت. اول اینکه آیا سرپرست همچنان مشغول به تحصیل هست؟ و سؤال دیگر اینکه مدرک تحصیلی سرپرست چیست؟ (این سؤال با اعداد ۱ تا ۹ پاسخ داده شده است). من تصور می کنم با جایگذاری ۰ به جای متغیرهای تهی موجود در ویژگی مدرک تحصیلی هم به سؤال اول (آیا سرپرست با سواد است؟) پاسخ دادیم، و هم می توان ویژگی مربوط به ستون اول را به کلی حذف کرد.

مورد دوم اینکه بدیهی است که افرادی که سواد ندارند مشغول به تحصیل هم نیستند. (در اینجا رسته مشغول به تحصیل نیستند با ۱ و مشغولین به تحصیل را با ۰ نمایش داده شده است).

```

1 #deal with NAs
2 > X2[is.na(X2[,8]),8]<-0 #Degree of
   education
3 > X2[is.na(X2[,7]),7]<-1 #Study employment
4 > X2<-X2[, -6]
5 > dim(X2)
6 [1] 2331    61
7 > dim(na.omit(X2))
8 [1] 2331    61
9 > colnames(X2)[1]<-"shahr" #This command is
   only for correcting the name of the
   first column and does not have much
   point

```

## ۳-۴-۱ افزار داده ها

در این بخش می خواهم داده ها را به سه بخش مجموعه آموزشی<sup>۱۸</sup>، مجموعه اعتبار سنجی<sup>۱۹</sup> و مجموعه آزمون<sup>۲۰</sup> افزار<sup>۲۱</sup> کنیم. میدانیم که بیشترین سهم مربوط به مجموعه آموزشی است لذا ۶۰ درصد داده ها را به صورت تصادفی<sup>۲۲</sup> برای این مجموعه انتخاب می کنیم. ۳۰ درصد داده ها برای مجموعه اعتبار سنجی و ۱۰ درصد برای مجموعه آزمون به تصادف انتخاب می شود. اجرای کد پیش رو این فرایند را برای ما اجرا می کند.

```
1 X2<-cbind(X2,Y_classification)
```

<sup>16</sup> Null

<sup>17</sup> Missing Value

<sup>18</sup> Training Set

<sup>19</sup> Validation Set

<sup>20</sup> Test Set

<sup>21</sup> Partition

<sup>22</sup> Random

دو مقدار اتخاذ می کردن) ولیکن این اعداد ۰ و ۱ نبودند. به عنوان مثال ۱ و ۲ یا ۱۳ و ۱۴ را برای پاسخ به یک سؤال صحیح یا غلط انتخاب می کردند. من با تعریف تابع<sup>۱۰</sup> بدون کسر از کلیت فلسفی در ماتریس داده ها، درایه های متناظر با این ویژگی ها را به رسته های دودویی ۰ و ۱ تبدیل کردم. نکته حائز اهمیت این است پس از اعمال تغییرات ممکن است برخی پیشگوها منطقی به نظر نرسند. اما هیچ چیز از کلیت موضوع کاسته نشده است. به عنوان نمونه متغیر SarparstSavad پس از اعمال تغییرات با ۱ سرپرست های بی سواد و با ۰ سرپرست های با سواد را نمایش می دهد.

این تابع دو خصوصیت جالب دیگر دارد. اولاً برای کار با داده های جدید هر آنچه روی داده ها اعمال می کند را به عنوان خروجی در اختیار مان می گذارد. لذا اگر با داده جدید و خارج از مجموعه داده های در دسترس رو به رو شویم به راحتی می توانیم تغییرات را مجدداً اعمال کنیم. نکته دوم این تابع ویژگی های بی اثر را که فقط نمایانگر عددی ثابت هستند را حذف می کند.

```

1 #this is a function for converting two-batch
   variables to binary also it removes
   constant features( input should be a
   dataframe also output can help us to
   apply transforms these changes on the
   unspoilt new data)
2 #for example def_make_binary()
3 def_make_binary <- function(df){
4   A<-df
5   hist<-c()
6   c<-rep(0,dim(df)[2])
7   for (i in 1:dim(df)[2]){
8     if(length(unique(na.omit(df[,i]))))
9       ==2){
10       tmp<-as.numeric(df[,i])
11       hist[i]<-min(na.omit(tmp))
12       tmp<-tmp-min(na.omit(tmp))
13       tmp<-tmp/max(na.omit(tmp))
14       A[,i]<-tmp
15     }
16     if(length(unique(na.omit(df[,i])))
17     ==1){
18       c[i]<- 1
19     }
20   }
21
22 #Apply
23 Apply1<-def_make_binary(X)
24 X2<-Apply1[[1]]

```

<sup>15</sup> Function

```

18 [5] "HazineKhoraki"
19 [6] "HazineNoshidani"
20 [7] "HazinePoshak"
21 [8] "HazineMaskan"
22 [9] "HazineLavazemKhanegi"
23 [10] "HazineDarmani"
24 [11] "HazineHamlonagh1"
25 [12] "HazineErtebatat"
26 [13] "HazineTafrihatFarhangi"
27 [14] "HazineGhazaAmade"
28 [15] "HazineKalaMotefaregheh"
29 [16] "HazineKalaBadavam"

```

```

2 set.seed(1)
3 train.rows<-sample(rownames(X2),dim(X2)[1]*0.6)
4 valid.rows<-sample(setdiff(rownames(X2),train.rows),dim(X2)[1]*0.3)
5 test.rows<-setdiff(rownames(X2),union(train.rows,valid.rows))
6 training_set<-X2[train.rows,]
7 validation_set<-X2[valid.rows,]
8 test_set<-X2[test.rows,]

```

در انتهای برای گزارش از ماتریس نهایی داده‌ها گزارش زیر را مشاهده فرمایید.

#### ۵-۴-۱ نرمالیله کردن متغیرها

در اینجا از روش نرمالیله کردن داده‌ها با استفاده از Standard Scaler بهره خواهیم برد. [۵] با اجرای کد پیش رو ویژگی‌های کمی نرمالیله خواهند شد. ضمناً لازم به توضیح است که باقی متغیرها یا دودویی هستند و یا رسته‌ای بیش از دورسته، لذا نیازی به نرمالیله شدن ندارند.

این مرحله حتماً بایستی بعد از افزایش داده‌ها صورت بگیرد که اطلاعات مجموعه تست از طریق نرمالیله شدن به داخل مجموعه آموزشی اصطلاحاً نشست نکند و به تعبیر علمی، نشست داده<sup>۲۶</sup> رخ ندهد.

```

1 X_standard_scaler.train<-training_set
2 X_standard_scaler.valid<-validation_set
3 X_standard_scaler.test<-test_set
4 #-----#
5 for(i in c(3, 5, 11, 12, 50, 51, 52, 53, 54,
      55, 56, 57, 58, 59, 60, 61)){
6   X_standard_scaler.train[,i]=scale(
     training_set[i])
7 }
8 #-----#
9 for(j in c(3, 5, 11, 12, 50, 51, 52, 53, 54,
      55, 56, 57, 58, 59, 60, 61)){
10  for (i in 1:dim(validation_set)[1]){
11    X_standard_scaler.valid[i,j]=((
12      validation_set[i,j]-mean(training_set[,j]))/sd(training_set[,j])
13    )
14  }
15 #-----#
16 for(j in c(3, 5, 11, 12, 50, 51, 52, 53, 54,
      55, 56, 57, 58, 59, 60, 61)){
17  for (i in 1:dim(test_set)[1]){
18    X_standard_scaler.test[i,j]=((
19      test_set[i,j]-mean(training_set[,j]))/sd(
20      training_set[,j]))
21  }
22 }
23 
```

```

1 > dim(training_set)
2 [1] 1398 62
3 > dim(validation_set)
4 [1] 699 62
5 > dim(test_set)
6 [1] 234 62

```

#### ۴-۴-۱ دسته‌بندی ویژگی‌ها

در داده‌های من ۳۹ ویژگی دودویی، ۶ متغیر رسته‌ای بیش از ۲ رسته و ۱۶ ویژگی کمی<sup>۲۳</sup> وجود دارد. در قسمت‌های قبل به مسائل مربوط به متغیرهای دودویی رسیدگی شد. ولیکن رسیدگی به برخی متغیرها مانند متغیر کمی (نرمالیله<sup>۲۴</sup> کردن داده‌ها) و یا متغیر رسته‌ای (کدگذاری داده‌ها<sup>۲۵</sup>) بایستی بعد از افزایش داده‌ها صرفاً بر اساس اطلاعات داده‌های مجموعه آموزشی صورت بگیرد. اما در این قسمت، هر یک از ویژگی‌های کمی و رسته‌ای بیش از دو رسته را تفکیک خواهیم کرد.

```

1 #categorical
2 > colnames(X2[,c(1,7,8,9,10,13)])
3 [1] "shahr"
4 [2] "SarparstMadrak"
5 [3] "SarparstFaaliat"
6 [4] "SarparstZanashoyi"
7 [5] "NahveTasarof"
8 [6] "NoeEskelet"
9
10
11 #numerical
12 > colnames(X2[,c(3,5,11,12,50,51,52,53,
13 ,54,55,56,57,58,59,60,61)])
14 [1] "TedadAza"
15 [2] "SarparstSen"
16 [3] "TedadOtagh"
17 [4] "SatheZirbana"

```

<sup>23</sup>Numerical

<sup>24</sup>Normalized

<sup>25</sup>Data Encoding

<sup>26</sup>Data Leakage

```

26 tmp[,1:6] <- tmp[,63:68]
27 tmp<-tmp[,1:62]
28 colnames(tmp)[1:6] <- tmp2
29 X_standard_scaler.valid<-tmp
30 rm(tmp,tmp2)
31
32 #apply on the test set
33 tmp<-target_encode( X_standard_scaler.test,
34   target_encoding = target_encod)
34 tmp<-tmp[,-63]
35 tmp2<-colnames(tmp)[63:68]
36 tmp[,1:6] <- tmp[,63:68]
37 tmp<-tmp[,1:62]
38 colnames(tmp)[1:6] <- tmp2
39 X_standard_scaler.test<-tmp
40 rm(tmp,tmp2)

```

با توجه به اینکه در فصل بعدی ممکن است مقادیر کمی منسوب به متغیرهای رسته‌ای من تغییر کند فعلاً در این بخش از واردکردن این مقادیر در ماتریس اصلی داده‌ها خودداری می‌کنم. در ادامه بعد از نهایی شدن این مقادیر با داده‌های اصلی آنها را جایگزین خواهم کرد. ضمناً با توجه به کافی بودن ثبت‌ها علیرغم نامتعادل<sup>۲۸</sup> بودن رده توفیق و رده عدم توفیق، نیازی به بیش نمونه گیری<sup>۲۹</sup> ندیدم چراکه تعادل<sup>۳۰</sup> رده توفیق در هر افزار من تقریباً ثابت است.

```

18      }
19 }

```

#### ۶-۴-۱ کدگذاری متغیرهای رسته‌ای

با توجه به اینکه مجموع تعداد رسته‌های متغیر رسته‌ای من در ویژگی‌های رسته‌ای مختلف عدد کوچکی نیست، استفاده از One hot encoding و Dummy encoding منجر به افزایش چشمگیر تعداد ویژگی‌ها می‌شوند. لذا بهتر است در این مورد از روش مرسوم Target encoding<sup>۲۷</sup> بر اساس میانه<sup>۲۸</sup> میانه<sup>۲۹</sup> متغیر پاسخ استفاده کنم. [۹] این مرحله نیز مانند بخش قبل حتماً باستگی پس از بخش افزایش داده‌ها انجام شود. دلیل این امر هم مانند آنچه در بخش قبل تشریح کردہ است.

```

1
2 #Start the target encoding process
3 for (i in c(1,7,8,9,10,13)){
4   X2[,i]<-as.factor(X2[,i])
5   X_standard_scaler.train[,i]<-as.factor(X_
6   _standard_scaler.train[,i])
7   X_standard_scaler.valid[,i]<-as.factor(X_
8   _standard_scaler.valid[,i])
9   X_standard_scaler.test[,i]<-as.factor(X_
10 _standard_scaler.test[,i])
11
12 #apply on the training set
13 tmp<-target_encode( X_standard_scaler.train,
14   target_encoding = target_encod)
14 tmp<-tmp[,-63]
15 tmp2<-colnames(tmp)[63:68]
16 tmp[,1:6] <- tmp[,63:68]
17 tmp<-tmp[,1:62]
18 colnames(tmp)[1:6] <- tmp2
19 X_standard_scaler.train<-tmp
20 rm(tmp,tmp2)
21
22 #apply on the validation set
23 tmp<-target_encode( X_standard_scaler.valid,
24   target_encoding = target_encod)
24 tmp<-tmp[,-63]
25 tmp2<-colnames(tmp)[63:68]

```

---

<sup>28</sup>Imbalance

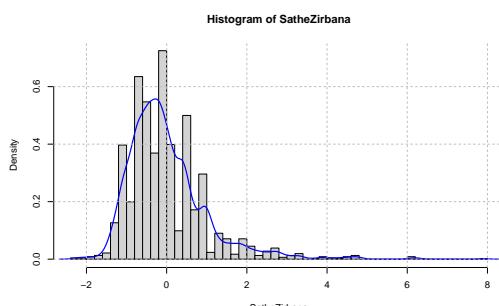
<sup>29</sup>Oversampling

<sup>30</sup>Balance

<sup>27</sup>Median

کم سن ترین سرپرست‌های خانوار سن‌هایی بین ۲۰ تا ۲۴ سال دارند. در مقابل مسن‌ترین سرپرست‌ها سن‌های از ۹۴ تا ۹۷ سال دارند. میانگین<sup>۳۷</sup> سن سرپرست‌های خانوارها کمتر از ۵۲ سال است و میانه این ویژگی عدد ۵۰ سال را نشان می‌دهد. گفته‌ی است که نیمی از سرپرست‌های خانوارهای موردمطالعه من‌سنی بین ۴۰ تا ۶۲ سال دارند. در مقابل تنها کمتر از ۵ درصد خانوارها سرپرستی با سن ۳۱ سال یا کمتر دارند.

نکته قابل توجه دیگر این است که همبستگی<sup>۳۸</sup> این متغیر با  $Y_{\text{classification}}$  که متغیر پاسخ<sup>۳۹</sup> من است برابر ۵۵٪ است. این در حالی است که همبستگی همین متغیر با  $Y_{\text{regression}}$  برابر ۰٪ است. در هر دو حالت تأثیر کم این متغیر را در متغیر پاسخ شاهد هستیم. اما خالی از لطف نیست کمی در خصوص تقاضت این دو عدد توضیح دهم. فارغ از این متغیر خاص من پس از بررسی‌هایی که روی ویژگی‌های این مجموعه داده داشتم متوجه شدم که اگر متغیر پاسخ  $Y_{\text{regression}}$  می‌بود، برخی از ویژگی‌ها دیگری می‌شدند و حالا که  $Y_{\text{classification}}$  ویژگی‌های دیگری مؤثر هستند. این به این معنی است که در تشخیص اینکه یک خانوار به طور کل چه درآمدی دارد یک سری ویژگی مهم است و اینکه آیا آن یک خانوار در دهک برتر اقتصادی هست یا خیر ویژگی‌هایی نه لزوماً برابر با ویژگی‌های قبل حائز اهمیت هستند. در انتهای این بخش یک بخش تحت عنوان انتخاب ویژگی<sup>۴۰</sup> قرار خواهد داد و از زبان پایتون<sup>۴۱</sup> در این زمینه کمک کوچکی خواهد گرفت.



شکل ۲: بافت‌نگار ویژگی سطح زیربنای محل سکونت خانوار

شکل ۲:

این ویژگی، ویژگی‌ای است که من تصور می‌کنم ارتباط قابل قبولی با درآمد خانوار داشته باشد. در این ویژگی نیز مانند سایر

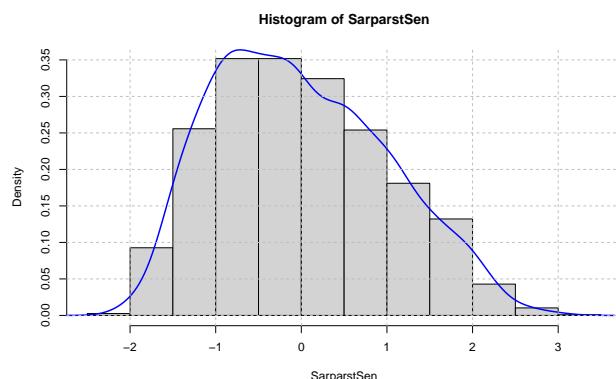
## ۲ تصویری سازی و اکتشاف داده‌ها

در فصل تصویری سازی احتیاج دارم که فرایند تصویری سازی روی کل داده‌ها صورت بگیرد. لذا افزایش‌های انجام شده در این بخش کاربردی ندارند.

### ۱-۲ بافت‌نگار ویژگی‌ها

در این بخش ویژگی‌هایی که قابلیت دارند از آنها بافت‌نگار<sup>۳۱</sup> رسم کنم را مورد تحلیل و بررسی قرار می‌دهم. این فرایند به من کمک می‌کند تا در مورد توزیع<sup>۳۲</sup> هر یک از ویژگی‌ها اطلاعات خوبی را به دست آورم.

در خصوص بازه‌های بافت‌نگار ویژگی‌ها، باید بگوییم حالت‌های مختلفی برای تعیین بازه‌های دسته‌بندی وجود دارد. هرچه طول بازه‌ها بیشتر باشد، نوشه<sup>۳۳</sup> ناشی از نمونه‌گیری تصادفی را کمتر به نمایش می‌گذارد. از طرف دیگر هرچه طول بازه‌ها کمتر باشد، تخمین بهتری از توزیع می‌توان پیدا کرد. [۷] من در این پژوهه تلاش کرده‌ام که طول بازه‌های دسته‌بندی را برای هر ویژگی به گونه‌ای تعیین شوند که بهترین ارتباط بصری را با مخاطب برقرار کنند. البته طول و تعداد بازه‌های انتخاب شده لزوماً بهترین نیستند و قضاوت این موضوع به عهده استاد محترم درس خواهد بود.



شکل ۱: بافت‌نگار سن سرپرست خانوار

شکل ۱:

قبل از رسم این بافت‌نگار تصویر من یک توزیع شبیه به توزیع نرمال<sup>۳۴</sup> بود. البته خیلی هم از آنچه تصویر می‌کردم فاصله ندارد و لیکن به وضوح یک چولگی<sup>۳۵</sup> کمی به سمت چپ وجود دارد و یا به اصطلاح علمی دقیق‌تر یک توزیع با چولگی مثبت<sup>۳۶</sup> است.

<sup>31</sup>Histogram

<sup>32</sup>Distribution

<sup>33</sup>Noise

<sup>34</sup>Normal Distribution

<sup>35</sup>Skewness

<sup>36</sup>Positive Skewness

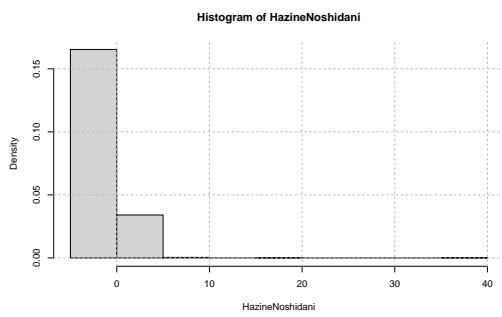
<sup>37</sup>Mean

<sup>38</sup>Correlation

<sup>39</sup>Response Variable

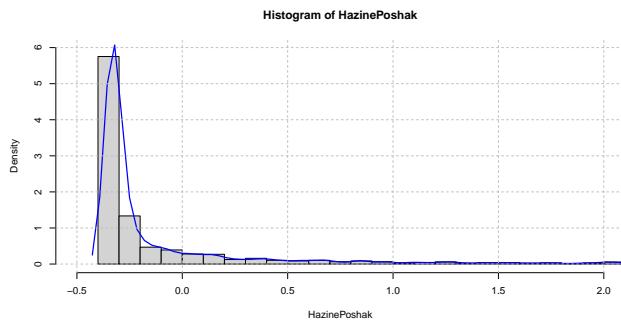
<sup>40</sup>Feature Selection

<sup>41</sup>Python

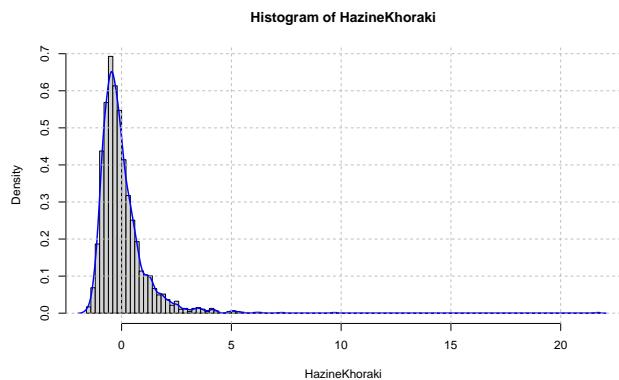


شکل ۴: بافت‌نگار ویژگی هزینه نوشیدنی

ویژگی‌های مربوط به ثروت در یک جامعه چولگی واضحی به سمت چپ را شاهد هستیم. میانه، عدد ۸۰ و میانگین عددی کمتر از ۸۶ مترمربع را نشان می‌دهد. ۵۰ درصد خانوارها در منزل‌هایی بین ۶۴ تا ۱۰۰ متر سکونت دارند و فقط ۵ درصد خانوارها در خانه‌هایی بیش از ۱۵۰ متر سکونت دارند. کمترین اعداد قابل مشاهده در این ویژگی خانه‌هایی با مساحت ۹ تا ۱۲ متر هستند و در مقابل بیشترین خانه‌های مشاهده شده در این مجموعه داده مساحتی بین ۲۵۰ تا ۳۶۰ متر مربع دارند. همبستگی این ویژگی با متغیر پاسخ  $\text{۰}^{۰}۷۷۵۳۰۵۸$  است لذا بین ۶۱ ویژگی می‌تواند از ویژگی‌هایی باشد که برای من حاوی اطلاعات مهمی است.



شکل ۵: بافت‌نگار ویژگی هزینه پوشاك



شکل ۳: بافت‌نگار ویژگی هزینه خوارک

شکل ۳: ویژگی هزینه خوارک از ویژگی‌هایی است که ممکن است در نگاه اول به نظر برسد در تعیین دهک برتر اقتصادی کمک خوبی به من نمی‌کند. این در حالی است که همبستگی  $\text{۰}^{۰}۳۲۷۷۶۵۱$  این ویژگی با متغیر پاسخ تا حد خوبی این ادعا را رد می‌کند. لازم به ذکر است که عدد  $\text{۰}^{۰}۳۲۷۷۶۵۱$ ، نشان دهنده همبستگی بالایی نیست اما در مقیاس ۶۱ ویژگی دیگر می‌تواند از ویژگی‌هایی باشد که حاوی اطلاعات خوبی برای رده‌بندی من خواهد بود.

کمترین داده‌ها نمایانگر اعداد ۴۵ تا ۱۱۹ هزار تومان به ازای هرماه، و بزرگ‌ترین داده‌های این ویژگی از ۸ تا ۲۱ میلیون تومان را نشان می‌دهند. لازم به ذکر است که ۵۰ درصد جامعه آماری موردمطالعه من از ۸۴۴ هزار تومان تا ۱ میلیون و ۶۷۱ هزار تومان در سال ۱۳۹۹ بابت خوارک در هرماه هزینه کرده‌اند.

شکل ۴:

هزینه نوشیدنی چولگی بسیار شدیدی به سمت چپ دارد. از طرفی همبستگی بالایی با متغیر پاسخ ندارد. در بخش انتخاب ویژگی در مورد اینکه این ویژگی را وارد مدل خواهم کرد یا خیر بیشتر توضیح خواهم داد.

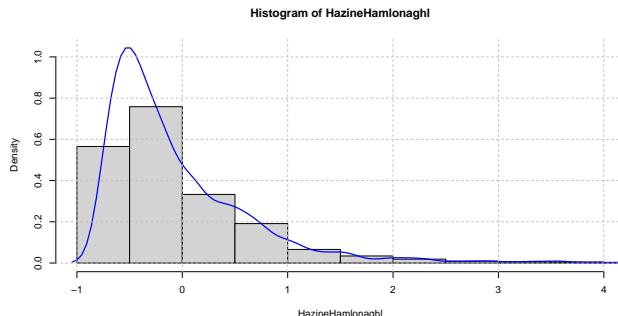
شکل ۵:

هزینه پوشاك با توجه به همبستگي  $\text{۰}^{۰}۲۱۶۳۸۴۶$  می‌تواند حاوی اطلاعات مفیدی برای من باشد. بیش از نیمی از جامعه مورد مطالعه هیچ هزینه‌ای بابت پوشاك در ماه قبل از پر کردن پرسشنامه نپرداخته‌اند. ۷۵ درصد از جامعه موردمطالعه کمتر از ماهی ۹۵ هزار تومان به طور متوسط هزینه پوشاك کرده‌اند و بیشترین هزینه در طی ماه عددی کمتر از ۷ میلیون و ۶۰۰ هزار تومان را نشان می‌دهد. این هزینه، شاید در مقابل هزینه‌های بزرگ‌تری مثل هزینه مسکن کم اهمیت تر باشد اما برآورد خوبی از استطاعت مالی افراد می‌دهد. چولگی این بافت‌نگار نیز به شکل واضحی به سمت چپ بوده اما حدس من این است که در رده‌بندی دهک برتر کارایی قابل قبولی خواهد داشت.

شکل ۶:

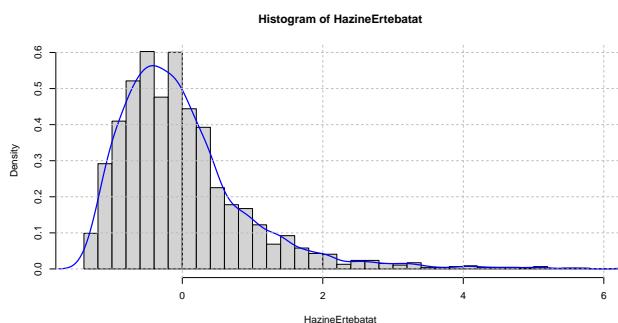
شاید از نظر شهودی یکی از ویژگی‌های بسیار مهم هزینه مسکن باشد اما از نظر همبستگی با متغیر پاسخ تقریباً عملکردی شبیه به ویژگی هزینه پوشاك را دارد. کمترین هزینه مسکن حدود ۱۱۶ هزار تومان و بیشترین هزینه مسکن بیش از ۸۰ میلیون تومان در هر ماه ثبت شده است. نیمی از خانوارها کمتر از ۲ میلیون و ۱۰۰ هزار تومان هزینه مسکن در هر ماه پرداخت کرده‌اند. همچنین ۷۵ درصد ثبت‌ها کمتر از ۳ میلیون و ۳۴۵ هزار تومان بابت هزینه مسکن متحمل هزینه شده‌اند. همبستگی این ویژگی با متغیر پاسخ  $\text{۰}^{۰}۲۵۸۱۶۸۵$  بوده و با ویژگی هزینه پوشاك  $\text{۰}^{۰}۹۲۴۴۹۷۷$  همبستگی

## می‌دهد هزینه‌های درمانی ارتباط خیلی قوی‌ای با دهک‌بندی اقتصادی خانوارها ندارد



شکل ۹: بافت‌نگار ویژگی هزینه‌های حمل و نقل

شکل ۹: ویژگی هزینه حمل و نقل همبستگی نسبی مناسبی با متغیر پاسخ دارد. این عدد  $0.2577946\%$  است و باید بگوییم شاخص‌های کمینه، چارک اول<sup>۴۲</sup>، میانگین، چارک سوم<sup>۴۳</sup> و بیشینه برای این ویژگی به ترتیب اعداد  $0, 745, 000, 1, 650, 000, 2, 491, 068$  و  $3, 345, 000$  و نهایتاً  $80, 500, 000$  بر حسب ریال در هر ماه می‌باشند.



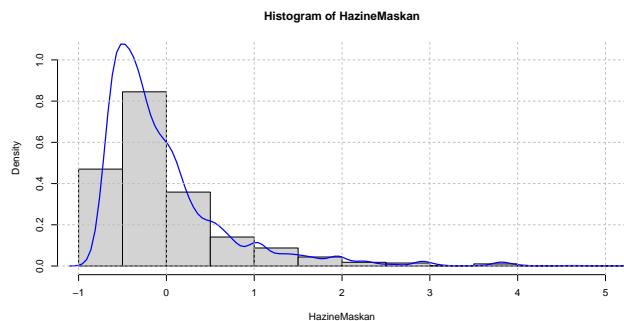
شکل ۱۰: بافت‌نگار ویژگی هزینه‌های ارتباطات

شکل ۱۰: هزینه‌های مربوط به بحث ارتباطات از جمله ویژگی‌هایی است که نه تنها از نظر بصری به ما اطلاعات قابل درکی می‌دهد بلکه همبستگی نسبتاً بالایی با متغیر پاسخ دارد. همبستگی این ویژگی با متغیر پاسخ حدود  $30.47933\%$  بوده و جدول شاخص‌های مرکزی این ویژگی به شرح ذیل است.

```
1 > summary(X$HazineErtebatat)
2   Min. 1st Qu. Median      Mean
3     0    600000 1000000 1182938
4   3rd Qu.      Max.
5 1500000 10500000
```

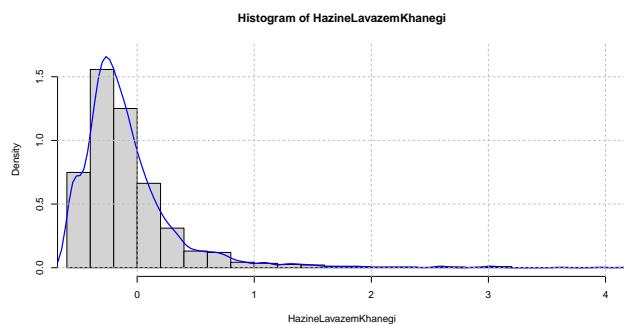
<sup>42</sup>First Quartile

<sup>43</sup>Third Quartile



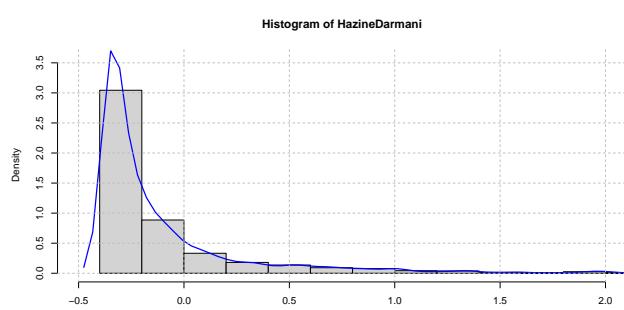
شکل ۶: بافت‌نگار ویژگی هزینه مسکن

دارد که نشان می‌دهد این دو ویژگی اطلاعات مجزایی رده‌بندی متغیر پاسخ در اختیار مان می‌گذارند.



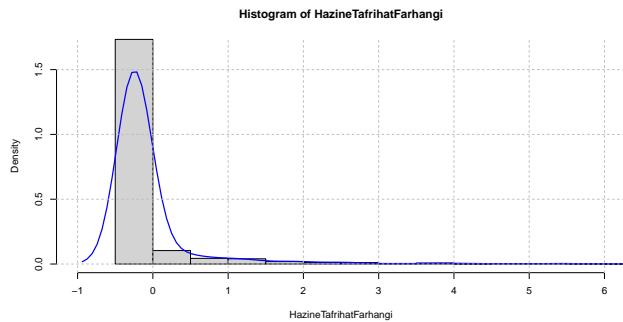
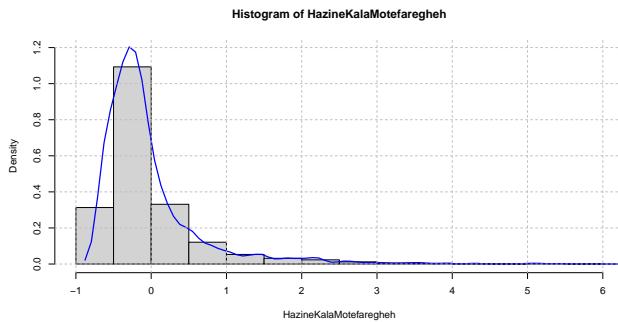
شکل ۷: بافت‌نگار ویژگی هزینه لوازم خانگی

شکل ۷: شکل بافت‌نگار هزینه لوازم خانگی شبیه به باقی هزینه‌های سنتی است اما با این حال تقریباً نصف ویژگی‌های دیگر با متغیر پاسخ همبستگی دارد. دلیل این امر می‌تواند مربوط به خرید اقساطی باشد.



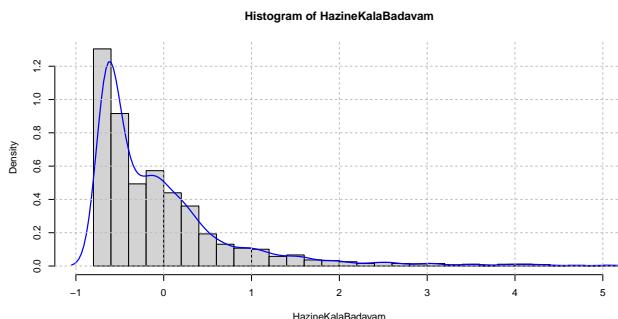
شکل ۸: بافت‌نگار ویژگی هزینه‌های درمانی

شکل ۸: شکل بافت‌نگار هزینه‌های درمانی نیز چولگی قابل توجهی به سمت چپ دارد اما همبستگی پایین آن با متغیر به پاسخ نشان



شکل ۱۳: بافت‌نگار ویژگی هزینه متفرقه

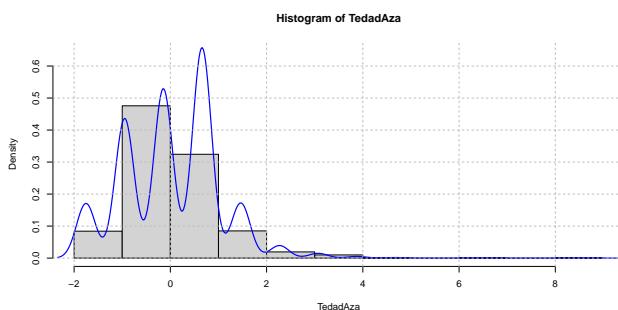
3	0	450000	800000	1196962
4	3rd Qu.	Max.		
5	1302000	74360000		



شکل ۱۴: بافت‌نگار ویژگی هزینه کالای بادوام

شکل ۱۴:

کالای بادوام به جهت گران قیمت بودن، معمولاً توسط افرادی که برتری اقتصادی دارند خریده می‌شود. در اینجا نیز بین این متغیر و متغیر پاسخ همبستگی  $0.3498422$  وجود دارد. لذا ویژگی حاوی اطلاعات نسبتاً ارزشمندی برای من خواهد بود.



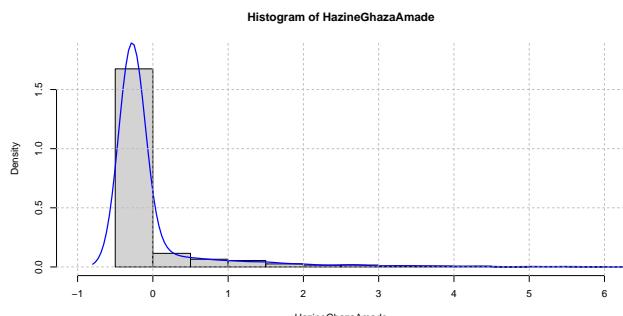
شکل ۱۵: بافت‌نگار تعداد اعضای خانوار

شکل ۱۵:

ویژگی تعداد اعضای خانوار نیز یک ویژگی کمی است که

شکل ۱۱: بافت‌نگار ویژگی هزینه تفریحات فرهنگی

شکل ۱۱: ویژگی هزینه تفریحات فرهنگی از ویژگی‌هایی است که همبستگی بسیار بالایی با  $Y_{regression}$  دارد اما متأسفانه همبستگی کمتری با متغیر پاسخ دارد. در هر حال می‌توان ادعا کرد که این ویژگی حاوی اطلاعات است به خصوص که سه چهارم ثبت‌ها هیچ هزینه‌ای در این خصوص انجام نداده‌اند.



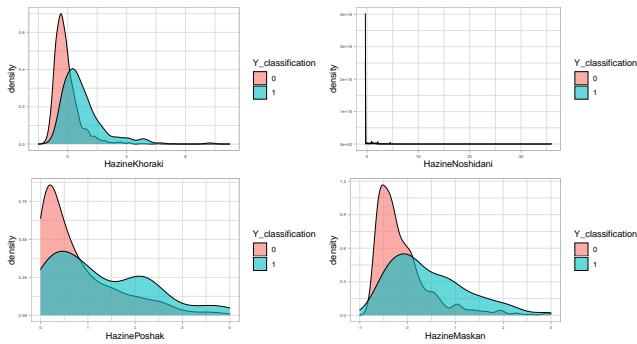
شکل ۱۲: بافت‌نگار ویژگی هزینه غذای آماده

شکل ۱۲: اساساً ویژگی‌هایی که در خصوص هزینه‌های لوكس هستند می‌توانند در رده‌بندی دهک برتر اقتصادی مفید باشند. این ویژگی نیز مانند هزینه‌های تفریحات فرهنگی در خصوص  $75$  درصد داده‌ها عدد صفر را نشان می‌دهد. این در حالی است که بیشترین رقم هزینه غذای آماده  $3$  میلیون تومان ثبت شده است. ضمناً همان‌طور که پیش‌بینی می‌شود، همبستگی نسبتاً خوبی با متغیر پاسخ دارد.

شکل ۱۴:

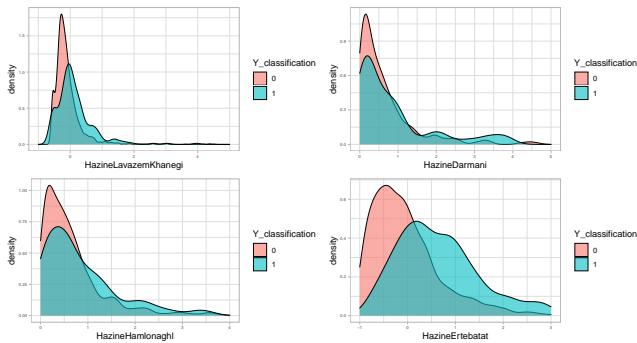
هزینه‌های متفرقه به گونه‌ای توزیع شده اند که بیشترین همبستگی را با  $Y_{regression}$  دارند. این نشان می‌دهد این ویژگی دارای اطلاعات ارزشمندی است. لذا بهتر است خروجی تابع `summary` را روی این ویژگی با هم بینیم.

```
1 > summary(X$HazineKalaMotefaregheh)
2   Min. 1st Qu. Median      Mean
```



شکل ۱۸: نمودار چگالی ویژگی‌های هزینه خوارکی، هزینه نوشیدنی، هزینه پوشак و هزینه مسکن

شکل ۱۸: در خصوص هزینه خوارکی و پوشاك می‌توان به محکمی ادعا کرد هزینه خوارکی خانوارهایی که در دهک برتر اقتصادی هستند به طور میانگین رقم بیشتری است. در خصوص هزینه پوشاك جالب است که رفتار افراد ۹ دهک اول با دهک دهم کاملاً متفاوت است. همچنین واضح است که خانوارهای دهک دهم، هزینه بیشتری بابت مسکن پرداخته‌اند.

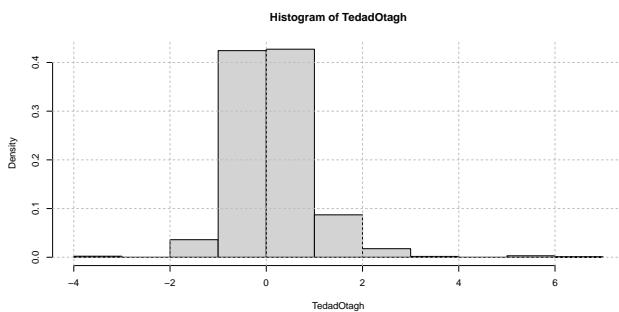


شکل ۱۹: نمودار چگالی ویژگی‌های هزینه لوازم خانگی، هزینه درمانی، هزینه حمل و نقل و هزینه ارتباطات

شکل ۱۹: به طور میانگین تفاوت چندانی میان خانوارهای دهک دهمی و سایر دهک‌ها در هزینه‌های درمانی و حمل و نقل وجود ندارد. البته کمی لوکس‌تر بودن را می‌توان در این ویژگی‌ها در مورد خانوارهای دهک دهمی مشاهده کرد اما تفاوت جدی‌ای وجود ندارد. در خصوص هزینه‌های لوازم خانگی خانوارهای دهک دهم، هزینه به نسبت بیشتری نسبت به خانوارهای دهک اول تا نهم پرداخته‌اند. اما در ویژگی هزینه ارتباطات تفاوت به مراتب مشهودتری میان خانوارهای دهک دهم و سایر خانوارها دیده می‌شود که این نکته می‌تواند حاوی اطلاعات مفیدی باشد.

شکل ۲۰: در هر چهار ویژگی اختلاف توان اقتصادی بین دهک‌ها به چشم

به صورت فوق توزیع شده است. این ویژگی با توجه به وابستگی نسبی به متغیر پاسخ می‌تواند تا حد کمی، حاوی اطلاعات باشد.

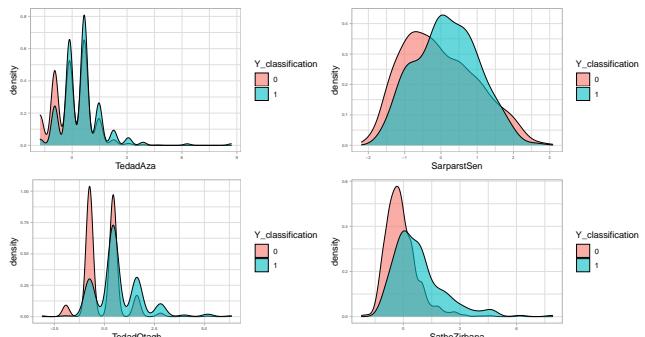


شکل ۱۶: بافت‌نگار تعداد اتفاق

شکل ۱۶: یکی از ویژگی‌های کمی بسیار مهم من، ویژگی تعداد اتفاق است که به صورت شهودی می‌توانم ادعا کنم با توان اقتصادی خانوارها بی‌ارتباط نیست. همبستگی نسبتاً بالای این ویژگی تأیید کننده ادعای من است.

اما آیا می‌توان فهمید که در این ویژگی‌های گفته شده، چه تفاوت محسوسی بین دهک برتر اقتصادی و سایر دهک‌ها وجود دارد؟ در بخش بعد که مربوط به نمودارهای چگالی<sup>۴۴</sup> است به این موضوع به طور جداگانه خواهیم پرداخت.

## ۲-۲ نمودار چگالی

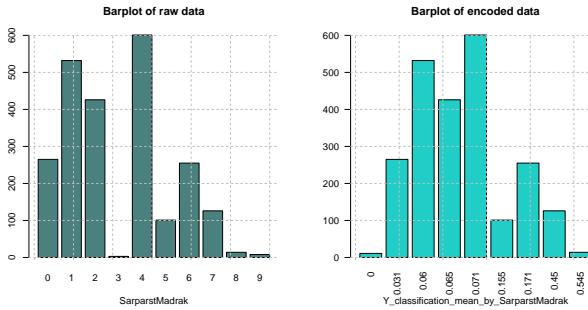


شکل ۱۷: نمودار چگالی ویژگی‌های تعداد اعضا، تعداد اتفاق، سن سرپرست و سطح زیربنا

شکل ۱۷: از این نمودارها می‌توان برداشت کرد که دهک برتر اقتصادی برخلاف تصور خانواده‌های پرجمعیت‌تری هستند. به علاوه در خانه‌های بزرگ‌تر با تعداد اتفاق‌های بیشتری زندگی می‌کنند. همچنین سن سرپرست خانوار به طور کلی در دهک برتر اقتصادی به طور میانگین بیشتر از سایر دهک‌هاست.

<sup>44</sup>Density

من تصور می‌کنم اطلاعات ارزشمندی برای ارائه در این ویژگی وجود نداشته باشد.

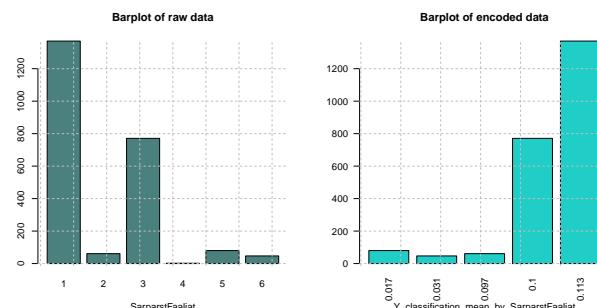


شکل ۲۲: نمودار میله‌ای ویژگی مدرک تحصیلی سرپرست خانوار

شکل ۲۲:

مدرک تحصیلی سرپرست خانوار (پس از کدگذاری) همبستگی خوبی با متغیر پاسخ دارد. این نشان می‌دهد این ویژگی می‌تواند حاوی اطلاعات مفید باشد. البته شهود شخصی من نیز با این تحلیل همخوانی دارد و از نظر من این نتیجه‌گیری دور از انتظار نبود.

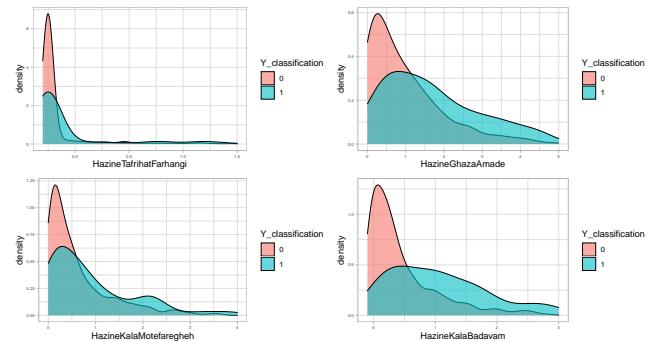
نکته مهم دیگر این است که در داده‌های خام (پس از جایگذاری به جای داده‌های تنهی (برای معرفی افراد بی‌سواد) من ۹ رسته برای توصیف این ویژگی داشتم. اما پس از کدگذاری، این ۹ رسته به ۸ رسته کاهش یافت. دلیل این امر این است که رسته‌های ۳ و ۹ هر دو پس از کدگذاری در یک رسته قرار گرفته‌اند. چرا که هیچ‌یک از داده‌هایی که این دو رسته را داشتند در دهک برتر اقتصادی قرار نگرفته‌اند.



شکل ۲۳: نمودار میله‌ای ویژگی فعالیت سرپرست خانوار

شکل ۲۳:

نکته قابل توجه در مورد این نمودار این است که در اینجا نیز شاهد کاهش نکته‌ای پس از کدگذاری هستیم. اما کاهش رسته‌های این ویژگی نکته‌ای جدید دارد. فقط در ۲ ثبت سرپرست خانوار محصل بوده است (rstه ۴). از آنجایی که کدگذاری بر

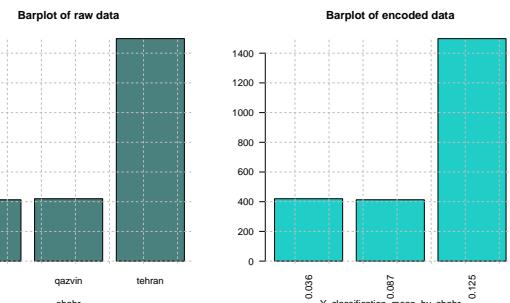


شکل ۲۰: نمودار چگالی ویژگی‌های هزینه تغیرات فرهنگی، هزینه غذای آماده، هزینه کالای متفرقه، هزینه کالای بادوام

می‌خورد. اما نکته‌ای که توجه من را به خود جلب کرده است این است که رفتار خانوارهای دهک‌های مختلف در هزینه تغیرات فرهنگی خیلی متفاوت است. همچنین خانوارهای پردرآمد به شکل قابل توجهی بیشتر از خانوارهای ۹ دهک اول هزینه‌های بیشتری بابت غذای آماده و کالای بادوام پرداخت می‌کنند.

## ۳-۲ نمودار میله‌ای

نمودار میله‌ای برای نمایش ویژگی‌های رسته‌ای مناسب است. از آنجایی که در ماتریس نهایی داده‌ها، من از نوعی کدگذاری استفاده کرده‌ام، در این قسمت نمودار میله‌ای داده‌های خام را نیز در کنار نمودارهای کدگذاری شده متناظر قرار خواهم داد.



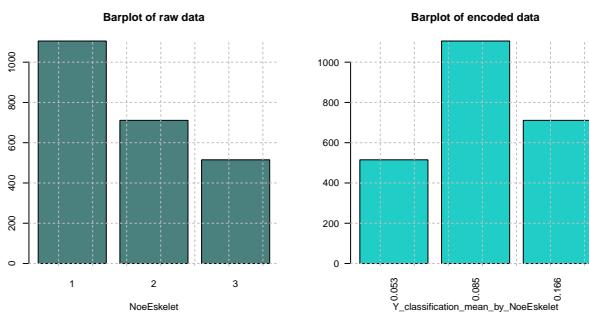
شکل ۲۱: نمودار میله‌ای ویژگی شهر (استان)

شکل ۲۱:

این نمودار میله‌ای به نوعی تصویری سازی جدول فراوانی<sup>۴۵</sup> داده‌ها از نظر شهر (استان) جمع‌آوری داده‌های است. تهران بیشترین فراوانی را دارد. ضمناً ترتیب میله‌ها در نمودار میله‌ای راست، با نمودار میله‌ای چپ یکی نیست. دلیل این امر نوع کدگذاری من است. این نکته در مورد همه نمودارهای میله‌ای این گزارش می‌تواند صادق باشد.

ضمناً این ویژگی همبستگی بسیار پایینی با متغیر پاسخ دارد و

<sup>45</sup>Frequency



شکل ۲۶: نمودار میله‌ای ویژگی نوع اسکت ساختمان تحت تصرف

شکل ۲۷:

در این شکل می‌توانید تمام نمودارهای میله‌ای متغیرهای رسته‌ای با رسته بیشتر از ۲ را در یک تصویر بینید. در این تصویر تلاش شده است تا توزیع ۰ و ۱ های متغیر پاسخ را در مورد هر رسته به صورت جداگانه به نمایش گذاشته شود.

### ۱-۳-۲ نمودار میله‌ای متغیرهای دودویی

متغیرهای دودویی می‌توانند متغیرهای رسته‌ای (با درسته) در نظر گرفته شوند. داده‌های من ۳۹ متغیر دودویی به عنوان پیشگو دارد. بررسی فراوانی ۰ و ۱ هر ویژگی اطلاعات خوبی به من می‌دهد. من برای ایجاد شهود بهتر، در هر نمودار اطلاعات اضافی ای درج کرده‌ام. لذا لازم است توضیح مختصری درخصوص نمودارها رائه کنم.

- در هر ویژگی، اطلاعات مربوط به برچسب ۰ سمت چپ و اطلاعات مربوط به برچسب ۱ سمت راست نمودار است.

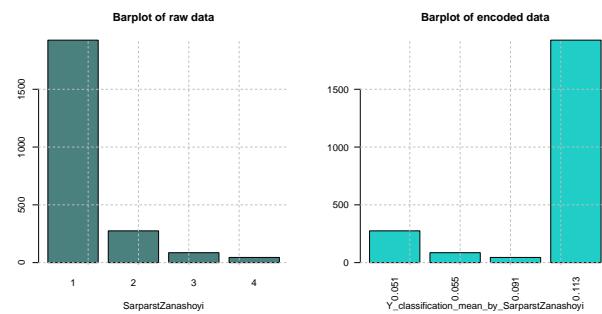
- به ازای هر برچسب، اعداد طبیعی ای روی میله‌های نمودار میله‌ای نوشته شده است که این اعداد به همراه رنگ‌بندی هر میله، نشان دهنده توزیع متغیر پاسخ در ثبت‌های مورد نظر است.

- یک عدد حقیقی<sup>۴۶</sup> در نیمه هر شکل (در راستای میله‌ها) قرار دارد که نشان می‌دهد چند درصد از ثبت‌های متغیر مذکور با ویژگی متناظر، در رده توفیق قرار دارد.

- اطلاعات مربوط به رنگ‌بندی در کل نمودارها یکسان بوده و در سمت راست هر نمودار مشخص شده است.

قبل از تصویری سازی بهتر است دقیقاً مشخصاً کنم چه چیزی را تصویری سازی می‌کنم. من در حقیقت جدول فراوانی هر متغیر را نسبت به متغیر پاسخ به تصویر می‌کشم. پیش از تصویری سازی

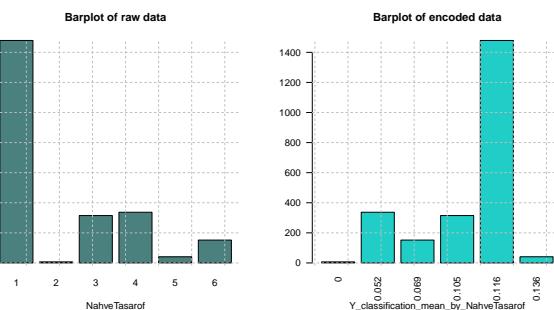
اساس میانگین متغیر پاسخ در مجموعه آموزشی انجام می‌شود، این دو ثابت در مجموعه آموزشی نبودند و در نتیجه کدگذاری ای برای آنها در نظر گرفته نشده است. اما هیچ جای نگرانی نیست در این موقع به جای متغیرهای جدید ۰ نسبت داده می‌شود. لذا در حقیقت تعداد رسته‌ها کم نشده است.



شکل ۲۴: نمودار میله‌ای ویژگی وضعیت زناشویی سرپرست خانوار

شکل ۲۴:

این ویژگی همبستگی بالایی با متغیر پاسخ ندارد و به نظر نمی‌رسد خیلی بتواند در فرایند رده‌بندی به من کمک کند. در انتهای این فصل در بخش انتخاب ویژگی به این موضوعات بیشتر خواهم پرداخت.



شکل ۲۵: نمودار میله‌ای ویژگی نحوه تصرف

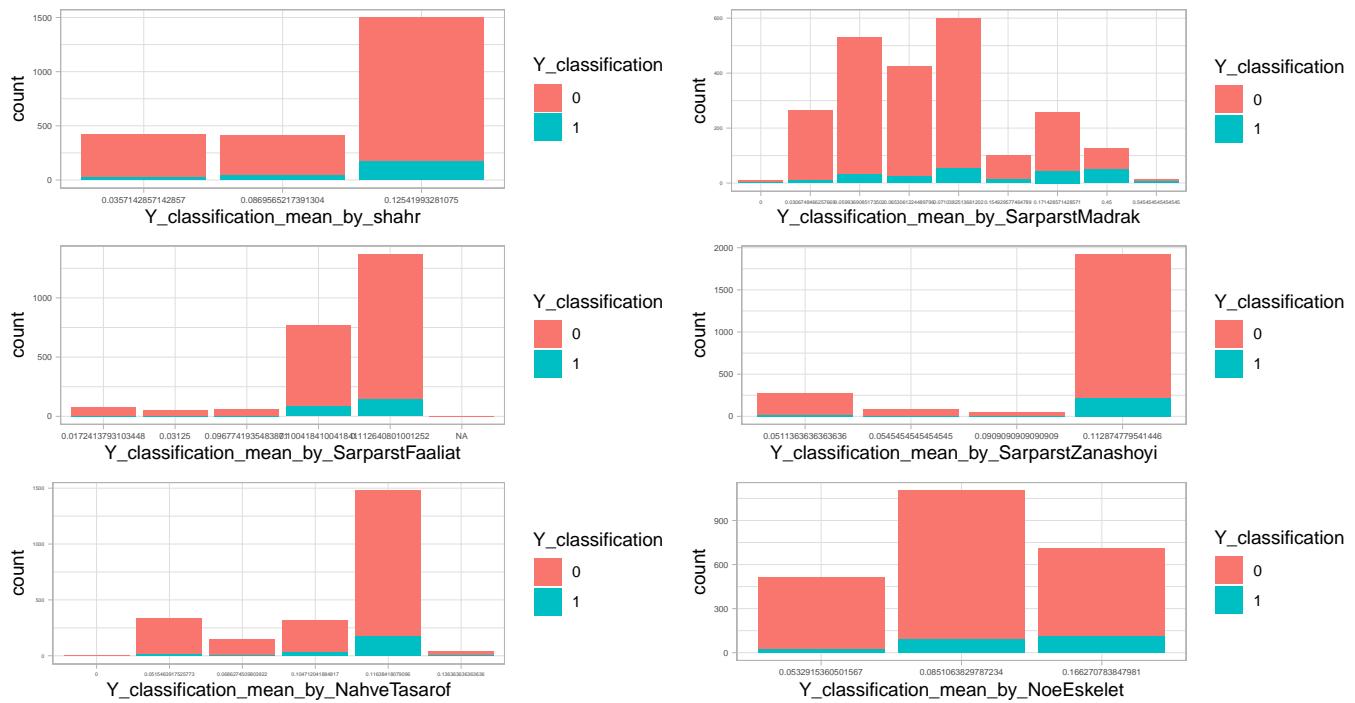
شکل ۲۵:

این نمودار نشان می‌دهد بخش قابل توجهی از خانوارها در مالک زمین و بنای منزل سکونتی خودشان ساکن هستند. ضمناً همبستگی پایینی بین این متغیر و متغیر پاسخ وجود دارد. نکته خاص دیگری در این ویژگی برای بیان کردن وجود ندارد.

شکل ۲۶:

به نظر می‌آید ویژگی‌هایی که تقاضت بین آنها منجر به هزینه قابل توجه برای خانوارها خواهد شد، اهمیت بیشتری دارند. این ویژگی نیز از این قاعده مستثنی نیست و از ویژگی‌های مهم ما به جهت همبستگی بالا با متغیر پاسخ، به شمار می‌رود.

<sup>46</sup>Real Number



شکل ۲۷: نمودار میله‌ای ویژگی‌های رسته‌ای بیش از دو متغیر در یک قاب

```

27 docharkhe    0     1
28      0 1907   200
29      1 191    33
30 # - - - - -
31          Y_classification
32 radio     0     1
33      0 2078   231
34      1 20    2
35 # - - - - -
36          Y_classification
37 radiozabt  0     1
38      0 1961   202
39      1 137    31
40 # - - - - -
41          Y_classification
42 tv       0     1
43      0 2092   233
44      1 6     0
45 # - - - - -
46          Y_classification
47 tvrangi   0     1
48      0 29     4
49      1 2069  229
50 # - - - - -
51          Y_classification
52 video     0     1
53      0 1457   138
54      1 641    95
55 # - - - - -

```

یک گزارش از جدول فراوانی همه متغیرها بر اساس متغیر پاسخ،  
قرار خواهم داد.

	Y_classification	
1	0	1
2	NoeKhanevar	0 2093 233
3		1 5 0
4	# - - - - -	
5		
6	Y_classification	
7	SarparstJensiat	0 1772 218
8		1 326 15
9	# - - - - -	
10		
11	Y_classification	
12	SarparstTahsil	0 20 7
13		1 2078 226
14	# - - - - -	
15		
16	Y_classification	
17	mashin	0 1052 35
18		1 1046 198
19	# - - - - -	
20		
21	Y_classification	
22	motor	0 1902 197
23		1 196 36
24	# - - - - -	
25		
26	Y_classification	

```

109          1   30   1
110 # - - - - -
111          Y_classification
112 coolergazimoteharek    0   1
113          0 2093  230
114          1   5   3
115 # - - - - -
116          Y_classification
117 zarfshoyi    0   1
118          0 1934  144
119          1 164   89
120 # - - - - -
121          Y_classification
122 microfer     0   1
123          0 1810  125
124          1 288   108
125 # - - - - -
126          Y_classification
127 gazlolekeshi 0   1
128          0   1   0
129          1 2097  233
130 # - - - - -
131          Y_classification
132 telephone    0   1
133          0 647   18
134          1 1451  215
135 # - - - - -
136          Y_classification
137 internet    0   1
138          0 485   6
139          1 1613  227
140 # - - - - -
141          Y_classification
142 hamam       0   1
143          0   3   0
144          1 2095  233
145 # - - - - -
146          Y_classification
147 ashpazkhane 0   1
148          0   8   0
149          1 2090  233
150 # - - - - -
151          Y_classification
152 coolerabisabet 0   1
153          0 225   17
154          1 1873  216
155 # - - - - -
156          Y_classification
157 borodatmarkazi 0   1
158          0 1988  209
159          1 110   24
160 # - - - - -
161          Y_classification

```

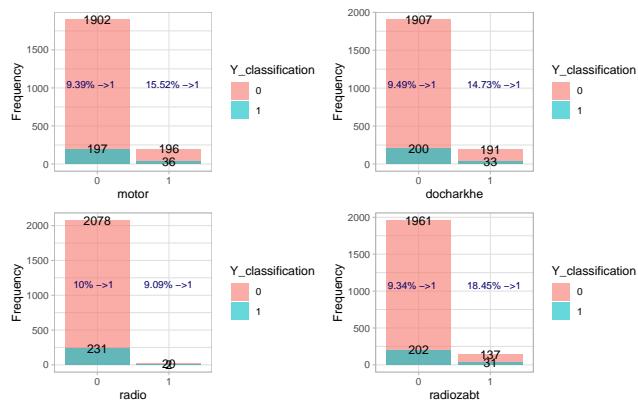
```

56          Y_classification
57 computer    0   1
58          0 1485  75
59          1 613   158
60 # - - - - -
61          Y_classification
62 mobile      0   1
63          0   70   6
64          1 2028  227
65 # - - - - -
66          Y_classification
67 freezer     0   1
68          0 1742  169
69          1 356   64
70 # - - - - -
71          Y_classification
72 yakhchhal   0   1
73          0 1594  169
74          1 504   64
75 # - - - - -
76          Y_classification
77 yakhchalfreezer 0   1
78          0 493   62
79          1 1605  171
80 # - - - - -
81          Y_classification
82 ojaghgaz    0   1
83          0   17   4
84          1 2081  229
85 # - - - - -
86          Y_classification
87 jarobarghi  0   1
88          0   87   5
89          1 2011  228
90 # - - - - -
91          Y_classification
92 lebasshoyi  0   1
93          0 183   7
94          1 1915  226
95 # - - - - -
96          Y_classification
97 khayati     0   1
98          0 1259  110
99          1 839   123
100 # - - - - -
101          Y_classification
102 panke       0   1
103          0 1933  198
104          1 165   35
105 # - - - - -
106          Y_classification
107 coolerabimoteharek 0   1
108          0 2068  232

```

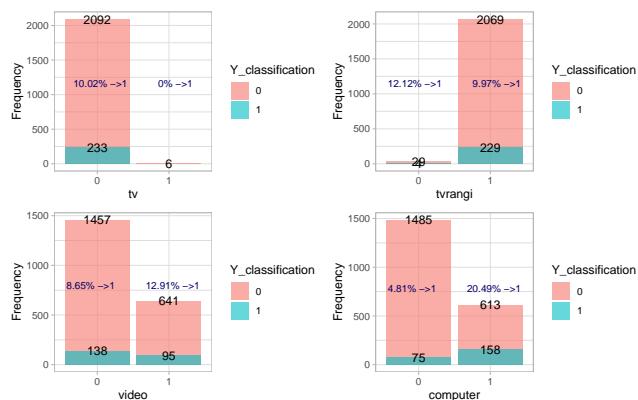
ویژگی‌های نوع خانوار و اشتغال به تحصیل سرپرست خانوار به علت نامتعادل بودن شدید اطلاعات خاصی به مدل‌های من نخواهند داد. (همان‌طور که قبلًاً توضیح داده‌ام لزوماً برچسب ۱ به معنای بله نیست)

در مقابل، ویژگی داشتن خودرو هم به دلیل تفاوت فاحش در توزیع ۰ و ۱ های متغیر پاسخ و همچنین تعادل مناسب بايستی اطلاعات خوبی در اختیار ما بگذارد. البته تفاوت در توزیع رده توفیق در خصوص جنسیت سرپرست نیز قطعاً حاوی اطلاعات مفید برای ردبندی خواهد بود.



شکل ۲۹: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن موتور، دوچرخه رادیو و رادیو ضبط

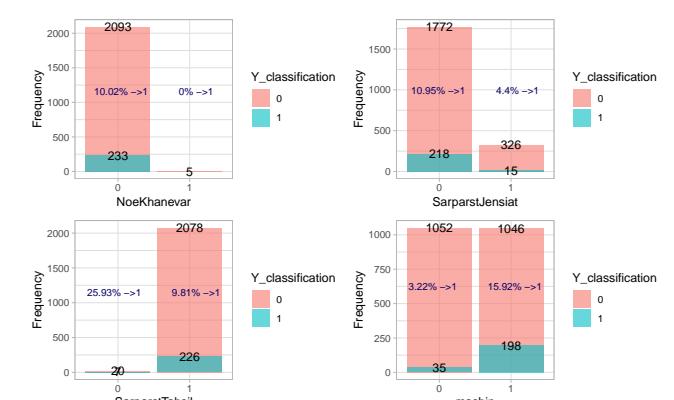
شکل ۲۹: به طور کلی این ویژگی‌ها نامتعادل هستند اما تعادل متغیر پاسخ در آنها (به استثنای ویژگی رادیو) در رسته‌های ۰ و ۱ متفاوت است. اینکه این ویژگی‌ها دارای اطلاعات هستند یا خیر در ادامه مشخص خواهد شد.



شکل ۳۰: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن تلویزیون، تلویزیون رنگی، کامپیوتر و دستگاه ویدیو

شکل ۳۰: این نمودارها به من کمک می‌کند تا برخی ویژگی‌هایی مثل

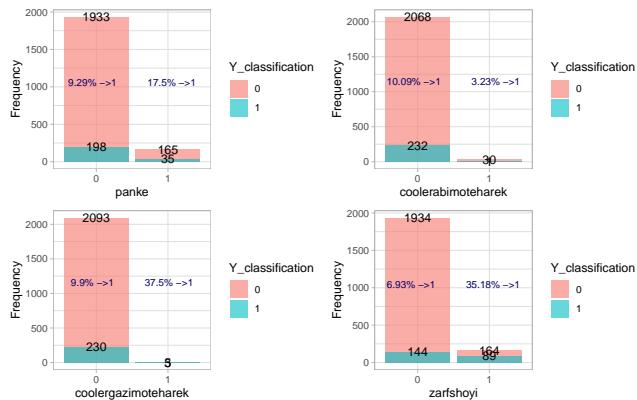
162	hararatmarkazi	0	1
163		0	1765 149
164		1	333 84
165	# -----		
166	Y_classification		
167	package	0	1
168		0	1677 162
169		1	421 71
170	# -----		
171	Y_classification		
172	coolergazisabet	0	1
173		0	2015 213
174		1	83 20
175	# -----		
176	Y_classification		
177	fazelabshahri	0	1
178		0	1162 84
179		1	936 149
180	# -----		
181	Y_classification		
182	nsokhtpokhtpaz	0	1
183		0	0 1
184		1	2098 232
185	# -----		
186	Y_classification		
187	nsokhtgarma	0	1
188		0	0 1
189		1	2098 232
190	# -----		
191	Y_classification		
192	nsokhtabgarm	0	1
193		0	0 1
194		1	2098 232
195	# -----		



شکل ۲۸: نمودار میله‌ای ویژگی‌های دودویی نوع خانوار، جنسیت سرپرست، اشتغال به تحصیل سرپرست و داشتن خودرو شخصی

شکل ۲۸:

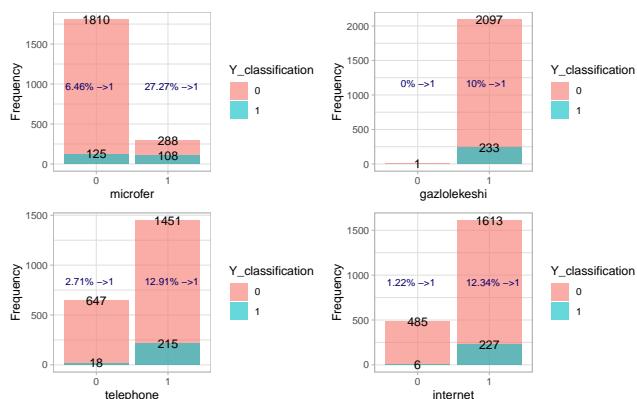
که نمی‌توانند حاوی اطلاعات مفیدی باشند. این نتایج نشان می‌دهد که نمودارهایی که در حال حاضر آن‌ها را بررسی می‌کنیم چقدر به کاهش بُعد من می‌تواند کمک کند. ضمناً تفاوت فاحش توزیع رده توفیق در ویژگی چرخ خیاطی نیز نکته حائز اهمیتی است که به نوبه خود می‌تواند حاوی اطلاعات خوبی باشد.



شکل ۳۳: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن پنکه، کولرگازی و آبی متحرک و ماشین ظرفشویی

شکل ۳۴:

ناکارآمد بودن ویژگی‌های مربوط به کولرهای متحرک کاملاً مشهود است. در مقابل تفاوت تعادل توزیع رده توفیق در ماشین ظرفشویی حرف‌های زیادی برای گفتن دارد.



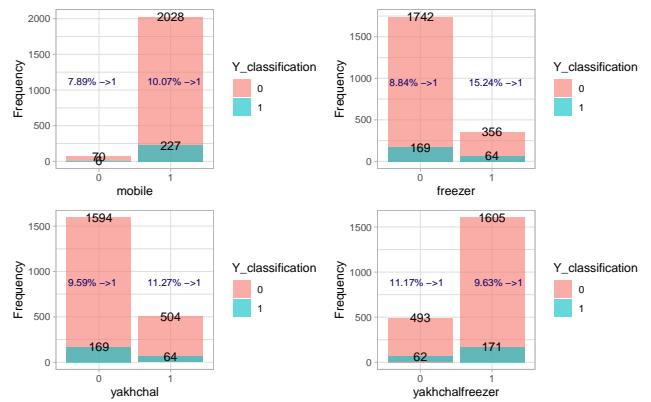
شکل ۳۴: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن مايكروويو، گاز لوله‌کشی، اينترنت و تلفن

شکل ۳۵:

من در فصل قبل ویژگی‌هایی که همگی یک عدد یا رسته را نمایش می‌دادند را حذف کردم. اما مثلاً در این جا و یا بسیاری از نمودارهای دیگر، ویژگی‌هایی مثل گاز لوله‌کشی وجود دارد که علماً با یک عدد ثابت تفاوتی ندارند. یکی از اقدامات ضروری من پس از این فصل بايستی حذف این متغیرها باشد. در مقابل ویژگی‌هایی مثل اينترنت و تلفن اگر رسته ۰ را نشان

تلويزيون رنگی و تلویزيون که تقریباً در همه داده‌ها یکسان است را شناسایی و آنها را حذف کنم. ضمناً ویژگی کامپیوتر با توجه به مواردی که در نمودارهای قبل گفته‌ام می‌تواند یک ویژگی مهم به شمار برود.

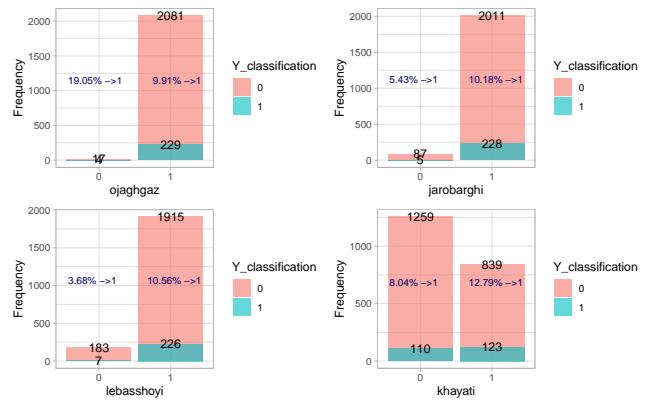
ضمناً بسیاری از ویژگی‌ها با یکدیگر هم‌پوشانی دارند. این اصلاً برای داده‌ها خوب نیست. به عنوان مثال ویژگی‌های رادیو و رادیوضبط، یا ویژگی‌های تلویزیون و تلویزیون رنگی (قبل تر هم ویژگی سواد و مدرک تحصیلی را ادغام کرده‌ایم) اساساً بهتر است پیشگووها مستقل از هم باشند.



شکل ۳۱: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن موبایل، فریزر، یخچال و یخچال فریزر

شکل ۳۱:

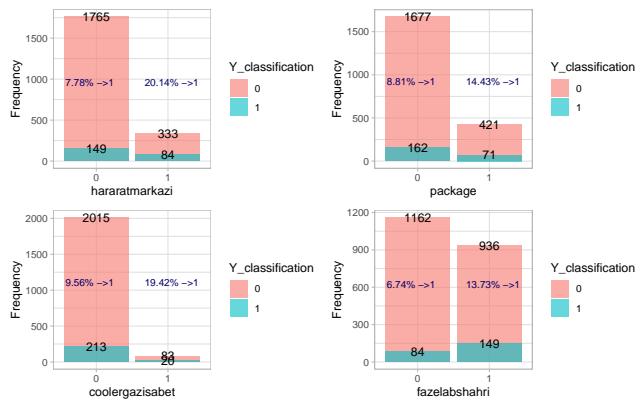
این تصویر هم حاوی نکته خاصی بیش از آنچه در نمودارهای قبل گفتم نیست فقط تها استدلال متفاوت من این است که خانوارهای دهک برتر اقتصادی علی‌الظاهر تمایل بیشتری برای خرید جدأگانه یخچال و فریزر دارند.



شکل ۳۲: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن اجاق گاز، ماشین لباسشویی، جاروبرقی و چرخ خیاطی

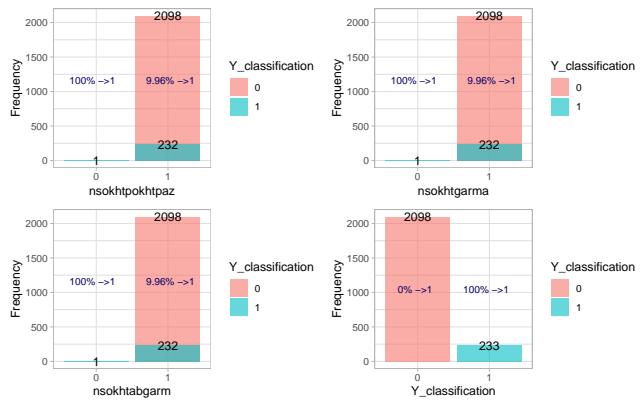
شکل ۳۲:

همچنان شاهد متغیرهایی هستیم که در حدی نامتعادل هستند



شکل ۳۶: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن سیستم حرارت مرکزی، پکیج، کولرگازی ثابت و فاضلاب شهری

این ویژگی می‌تواند یک ویژگی متمایزکننده خوبی باشد.



شکل ۳۷: نمودار میله‌ای ویژگی‌های دودویی انواع سوخت مصرفی برای مصارف مختلف و متغیر پاسخ

شکل ۳۷:

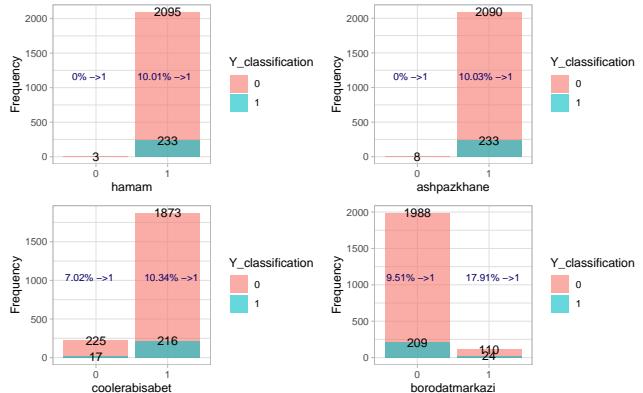
این نمودار هم ناکارآمدی ویژگی‌های مربوط به سوخت مصرفی را برای مصارف مختلف نشان می‌دهد. چراکه به‌جز یک ثبت (که در هر سه ویژگی مشترک‌اند) باقی ثبت‌ها از یک سوخت مشترک استفاده می‌کنند.

#### ۴-۲ نمودار جعبه‌ای

در این بخش تلاش خواهم کرد از نمودار جعبه‌ای<sup>۴۵</sup> و نمودار جعبه‌ای پهلو به پهلو<sup>۴۶</sup> برای نشان دادن توزیع ویژگی‌ها نسبت به متغیر پاسخ، استفاده کنم.

در مسائلی که متغیر پاسخ کمی یا به عبارت دیگر شبهی به خروجی رگرسیون<sup>۴۷</sup> باشد، نشان دادن ویژگی‌های دودویی مختلف در مقابل مقادیر متغیر پاسخ می‌تواند نمودارهای جعبه‌ای پهلو به

بدهند تقریباً می‌توان با احتمال بالایی ادعا کرد رده توفیق حاصل نشده است. لذا این ویژگی‌هایی که توزیع رده توفیق در یک سمت خیلی پایین است می‌توانند خیلی راهبردی باشند.



شکل ۳۵: نمودار میله‌ای ویژگی‌های دودویی داشتن یا نداشتن آسپزخانه، حمام، کولرآبی ثابت و سیستم برودت مرکزی

شکل ۳۵:

در اینجا نیز ویژگی‌های بدیهی‌ای مانند آسپزخانه و حمام برای ردبهندی وجود دارند که عدم تعادل شدید آنها نشان دهنده عدم کارایی آنها در ردبهندی خواهد بود. اما یک نکته جالب در خصوص سیستم برودت مرکزی و کولرآبی ثابت وجود دارد. تقریباً تعداد رده توفیق در خانوارهایی که سیستم برودت مرکزی دارند، با رده عدم توفیق خانوارهایی که کولرآبی ثابت ندارند تفاوت چندانی ندارد. لذا این حدس به ذهن من خطور کرد که بررسی کنم و ببینم چقدر این ثبت‌ها با یکدیگر اشتراک دارند. اما پس از بررسی متوجه شدم هیچ اشتراکی بین ثبت‌ها وجود ندارد. به عبارتی همه خانوارهایی که به سیستم برودت مرکزی دسترسی دارند، کولرآبی ثابت نیز دارند. قابل انکار نیست که این حدس به نتیجه قابل توجهی نرسید اما توجه به جزئیات مختلف می‌تواند در این فصل از یک گزارش داده‌کاوی بسیار کمک‌کننده باشد.

شکل ۳۶:

دقیقی که در تفسیر شکل ۳۵ به خرج دادیم اینجا کارساز شد. خانوارهایی که کولرگازی ثابت دارند و می‌توان آنها را در دهک برتر اقتصادی دانست، با کمی چشم‌پوشی از استثنای همان خانوارهایی هستند که در دهک برتر اقتصادی حضور داشتند ولی کولرآبی ثابت نداشتند. لذا درست است که ویژگی‌های مربوط به کولر خیلی نامتعادل هستند اما مثلاً می‌توان این نتیجه را گرفت که خانوارهایی که نه کولرآبی نه کولرگازی ثابت دارند احتمال بسیار بالا از ۹ دهک اول هستند. (که البته اکثر ثبت‌ها چنین شرایطی را ندارند)

نکته دیگر اینکه تفاوت تعادل خانوارهای دهک دهمی در دو رسته‌ی نمودار مربوط به ویژگی حرارت مرکزی مشهود است لذا

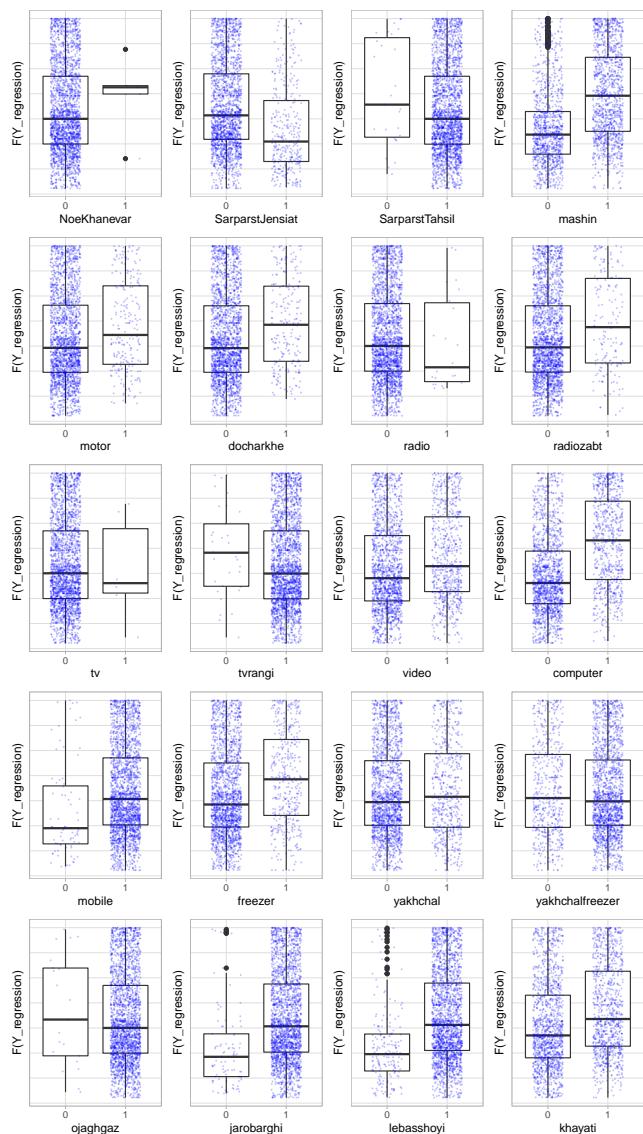
<sup>47</sup>Boxplot

<sup>48</sup>Side-By-Side Boxplot

<sup>49</sup>Regression

به طور متوسط می‌توان ادعا کرد در همه متغیرهای کمی تفاوت میان دهک برتر اقتصادی با دهک‌های دیگر قابل مشاهده است. شاید کمترین تفاوت بین این دو رده، در ویژگی سن سرپرست خانوار دیده شود. در مقابل در این نمودار بیشترین تفاوت مربوط به هزینه‌های تفریحات فرهنگی غذای آماده است. این ادعا نیز از طریق بررسی شکل ۲۰ نیز قابل تیجه‌گیری است.

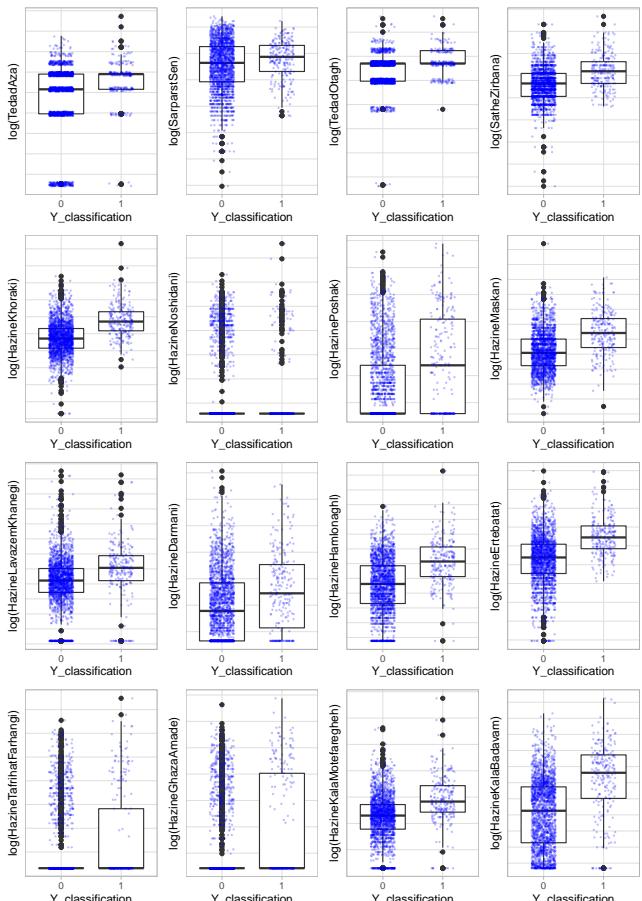
من در این نمودارها تلاش کرده‌ام که خود داده‌ها را نیز با یک تعداد نقطه‌آبی به عنوان نمونه من در ویژگی تعداد اتفاق که یک خود داده‌است به عنوان رفتار این ویژگی را که شبیه به یک متغیر رسته‌ای عمل کرده است را در تمایز با ویژگی‌ای مثل سن سرپرست خانوار شناسایی کنیم.



شکل ۳۹: نمودار جعبه‌ای پهلوی پهلو متغیرهای دودویی - تبدیلی صعودی از  $Y_{regression}$  از

پهلوی خوبی تولید کند. اما در این مسئله متغیر پاسخ من دودویی است. لذا به دو صورت برای به تصویر کشیدن داده‌ها می‌توان از نمودارهای جعبه‌ای پهلوی به پهلو استفاده کنم.

قبل از هر چیز باید بگوییم قرار دادن ویژگی‌های رسته‌ای در محورهای افقی و عمودی اطلاعات خوبی به من نمی‌دهند. لذا نمودارهای جعبه‌ای بهتر است برای متغیرهایی استفاده شود که محور عمودی متغیر کمی باشد.

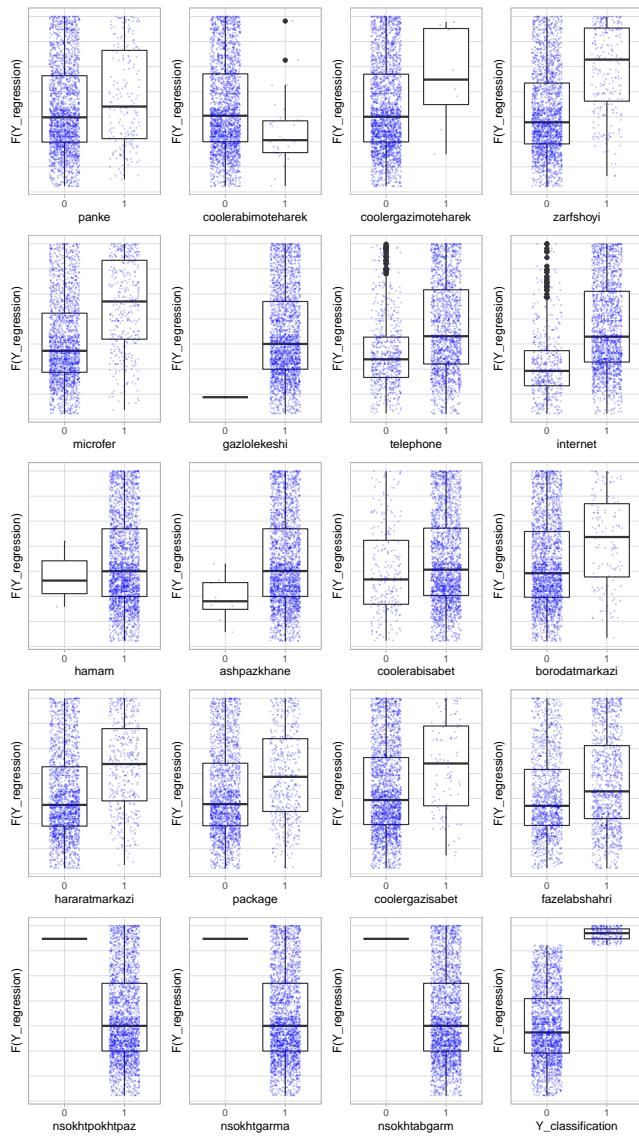


شکل ۳۸: نمودار جعبه‌ای پهلوی پهلو متغیرهای کمی - رده‌های پاسخ

شکل ۳۸:

استفاده اول من از نمودارهای جعبه‌ای پهلو به پهلو برای تولید نمودارهایی است که محور عمودی نمایانگر متغیرهای کمی مختلف باشد و در مقابل محور افقی نشان‌دهنده متغیر دودویی پاسخ باشد.

من برای ترسیم هر چه بهتر تغییرات در این نمودار به نوعی از لگاریتم<sup>۵۰</sup> متغیرهای کمی بهره برده‌ام. لذا بدون کاستن از کلیت، عددگذاری محور عمودی را حذف کردم که تمرکز فقط روی تغییرات باشد.



شکل ۴۰: نمودار جعبه‌ای پهلو به پهلو متغیرهای دودویی - تبدیل Y\_regression از صعده‌ی از

نظر بصری قابلیت درک بیشتری دارد. [۸]  
شکل ۴۱:

به طور کلی هر چه همبستگی پیشگوها با متغیر پاسخ بیشتر باشد و در مقابل، همبستگی پیشگوها با یکدیگر کمتر باشد آن پیشگو برای ما ارزشمندتر است. لذا پیشگو هایی که همبستگی ۱ با یکدیگر دارند حضور همزمانشان فقط به ابعاد اضافه می‌کند. لذا این نمودار می‌تواند یک مرجع بسیار خوب برای انتخاب ویژگی‌ها باشد.

چیزی که برای من جالب است این است که بدانم کدام ویژگی‌های من بیشترین همبستگی را با متغیر پاسخ دارند. البته برای دستیابی به نتایج دقیق‌تر بهتر است از خود R استفاده کنم. من از آوردن کد در این قسمت صرف نظر می‌کنم اما دوست دارم نتیجه متغیرهایی که بیشترین همبستگی را با متغیر پاسخ داشتند

استفاده دوم من از نمودارهای جعبه‌ای پهلو به پهلو برای تولید نمودارهایی است که محور عمودی آنها نمایانگر Y\_regression یا به عبارت دیگر یک تخمین خوب کمی از متغیر پاسخ است. محور افقی در این نمودارها مربوط به ویژگی‌های دودویی است. همچنین نقاطی برگرفته از خود داده‌ها برای تخمین چگالی بر روی نمودارها قرار داده شده است که تعادل ۰ و ۱ را مشخص می‌کند.

نکته اول این است که من برای نمایش بهتر Y\_regression یک تبدیل<sup>۵۱</sup> اکیداً صعده‌ی<sup>۵۲</sup> و غیرخطی اعمال کردام. دلیل این امر این بود که ساختهای میانه، و چارک<sup>۵۳</sup> ها بسیار به هم نزدیک می‌شوند و حتی با کمک تابع لگاریتم هم از هم فاصله نمی‌گرفتند. لذا به دو دلیل تحلیل عددی روی نمودار مذکور از اعتبار ساقط است. اولاً چون فاصله بین داده‌ها با توجه به تبدیل تغییر کرده و فقط ترتیب داده‌ها از نظر بزرگ و کوچک بودن Y\_regression حفظ شده است و لذا می‌توان مقایسه‌ای به شکل کمتر بودن یا بیشتر بودن داشت. دوماً از آنجایی که Y\_regression دقیقی از متغیر پاسخ ارائه نمی‌کند به طور کلی نتیجه‌گیری معتری از این نمودار نخواهم داشت اما صرفاً جهت بررسی تفاوت ۰ و ۱ ویژگی‌ها در Y\_regression می‌توان این نمودار را بررسی کرد.

شکل ۴۰:

در واقع این شکل و شکل قبل به نوعی زاویه دیدی دیگر برای توجیه آن چیزی است که از نمودارهای میله‌ای تحلیل کردام. اما در خصوص این شکل یک نمودار وجود دارد که محور عمودی Y\_regression و محور افقی Y\_classification با نگاهی اجمالی به این نمودار می‌توان فهمید که نه تنها تبدیلی که برای یکنواخت سازی Y\_regression انجام شده است، تبدیل قابل توجیهی است، بلکه این نمودار صحه‌ای مجدد بر دهک بندی درستی است که روی Y\_regression انجام شده است.

## ۵-۲ نمودار حرارتی

نمودار حرارتی<sup>۵۴</sup> یک نوع تصویرسازی داده‌ها<sup>۵۵</sup> است که در آن همبستگی بین ویژگی‌ها با یک رنگ نمایش داده‌می‌شود. ویژگی‌هایی که دارای همبستگی بالایی با یکدیگر هستند با رنگ‌های تیره‌تری نسبت به ویژگی‌هایی که دارای همبستگی کمتری نسبت به هم هستند به نمایش گذاشته می‌شوند. به طور معمول برای نمایش بصری اعداد یک بازه رنگی از روشن تا تیره (از کم به زیاد) در نظر گرفته می‌شود. البته این نمودار حرارتی از توسط یک ماتریس عددی قابل بیان هستند. اما نمودار حرارتی از

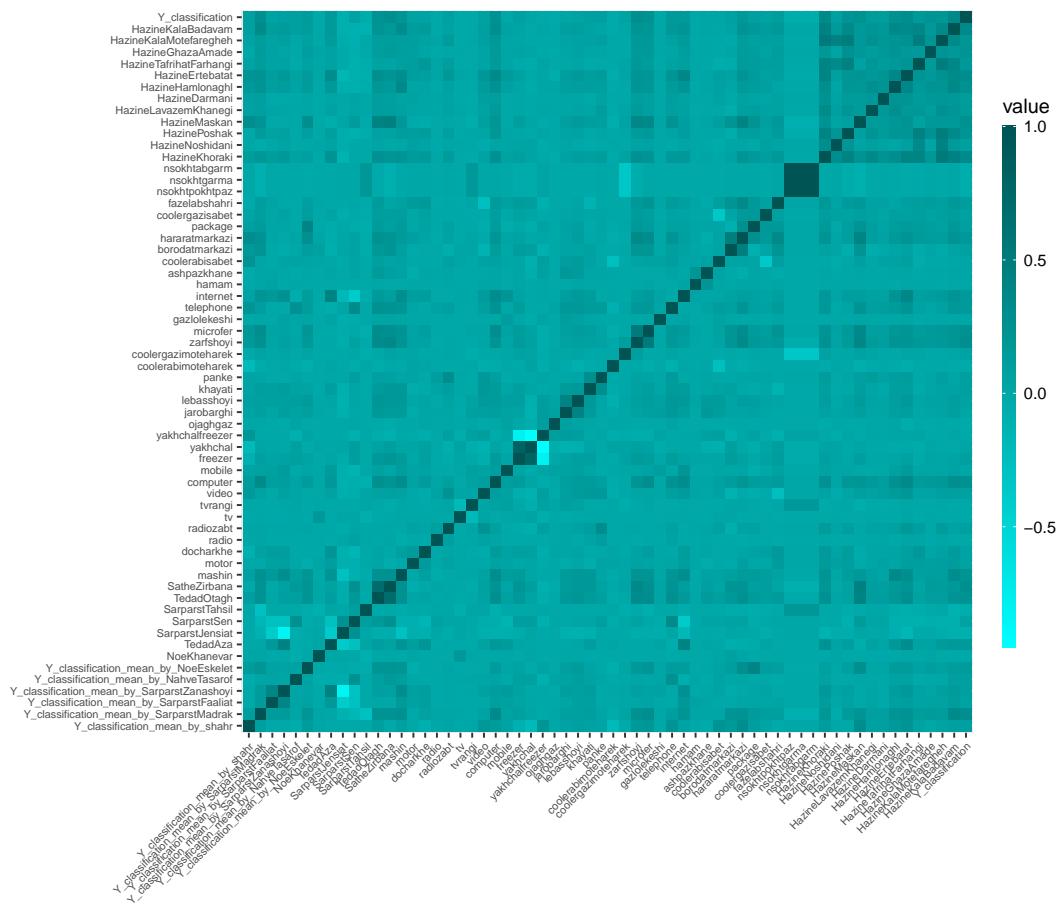
<sup>51</sup> Transformation

<sup>52</sup> Strictly Increasing

<sup>53</sup> Quartile

<sup>54</sup> Heatmap

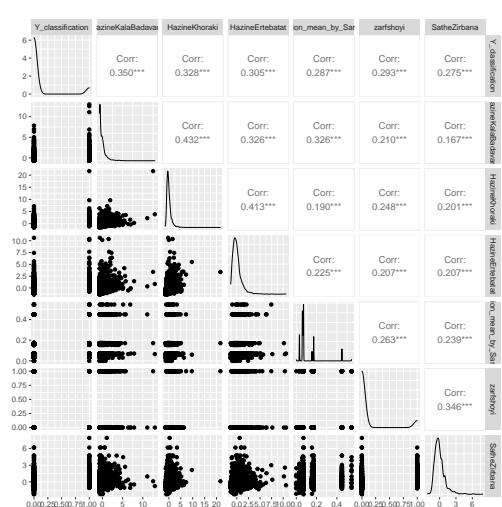
<sup>55</sup> Data Visualization



شکل ۴۱: نمودار حرارتی جدول همبستگی ویژگی ها

اختیار مدل های من خواهند گذاشت.

را در یک جدول به نمایش بگذارم.



شکل ۴۲: ماتریس نمودار پراکنش ویژگی های مهم

Var¹	Var²	value
۳۸۴۴ Y_classification	Y_classification	۰۰۰
۳۸۴۳ HazineKalaBadavam	Y_classification	۰۳۵
۳۸۴۲ HazineKhoraki	Y_classification	۰۳۳
۳۸۴۹ HazineErtebat	Y_classification	۰۳۰
۳۷۸۴ Y_classification_mean_by_SarparsTahlisi	Y_classification	۰۲۹
۳۸۱۶ zarfshoyi	Y_classification	۰۲۹
۳۷۹۵ SatheZirvana	Y_classification	۰۲۸

جدول ۱: جدول همبستگی، همسایه ترین متغیرها با متغیر پاسخ

از این رو بهتر است برای درک بهتر یک گروه از نمودار های پراکنش از ویژگی های مذکور بینیم. این کار به من کمک می کند تا درک بهتری از ارتباط این ویژگی هایی که به ظاهر مهم هستند در ذهن خود داشته باشم. از این رو شکل ۴۱ را ترسیم کردم در این شکل ویژگی هایی که در جدول ۵-۲ نمایش داده شده است، به صورت ماتریسی به تصویر کشیده شده.

نکته جالب در خصوص این نمودار این است که خود ویژگی ها با یکدیگر همبستگی خیلی بالایی ندارند و می توان به محکمی ادعا کرد هر یک از این ویژگی ها اطلاعات قابل توجه و مستقلی در

## ۶-۲ نمودار پراکنش سه بعدی

نشان می دهد که ویژگی های مهم ماتا حدی به داده های دهک برتر اقتصادی حساس هستند. همان طور که پیش تر گفتم روی داده ها یک تبدیل اکیداً صعودی اعمال شده است لذا مدل های من واقعاً با این داده های درون نمودار سروکار نخواهند داشت.

### ۷-۲ کاهش بعد

در یادگیری ماشین<sup>۵۹</sup> و آمار کاهش بعد<sup>۶۰</sup> یا کاهش ابعاد روند کاهش تعداد متغیرهای تصادفی راهنماییده<sup>۶۱</sup> [۴۹] از طریق به دست آوردن یک مجموعه از متغیرهای اصلی می باشد. کاهش ابعاد را می توان به انتخاب ویژگی و استخراج ویژگی تقسیم کرد.  
[۱۰]

من در این بخش مرکز خود را روی انتخاب ویژگی خواهم گذاشت. استثناناً در این بخش از زبان برنامه نویسی پایتون کمک کوچکی خواهم گرفت و مجدداً ادامه روند را به کمک زبان برنامه نویسی R ادامه خواهم داد.

من قبل از انجام این پروژه نتایج فوق العاده خوبی از خروجی پسته قدرتمند featurewiz گرفته بودم. هرچند ممکن است استفاده از یک زبان برنامه نویسی دیگر کمی غیر حر斐 ای به نظر بررسد اما نتوانستم از خروجی بسیار خوبی این پسته عبور کنم و از آن بهره مند نشوم.

این پسته با تکیه بر روش های تصادفی و گرادیان بوسٹینگ<sup>۶۲</sup> با قطعه<sup>۶۳</sup> قطعه کردن داده ها و مقایسه کردن آنها با متغیر پاسخ مهم ترین ویژگی ها را به عنوان ویژگی های با اهمیت معروفی می کند. مزیت بسیار خوب استفاده از این پسته، کاهش هوشمندانه بعد می باشد. من توابع این پسته روی داده های مجموعه آموزشی اعمال کرم و نتایج را به شرح پیش رو به دست آوردم.

شکل ۴۴:

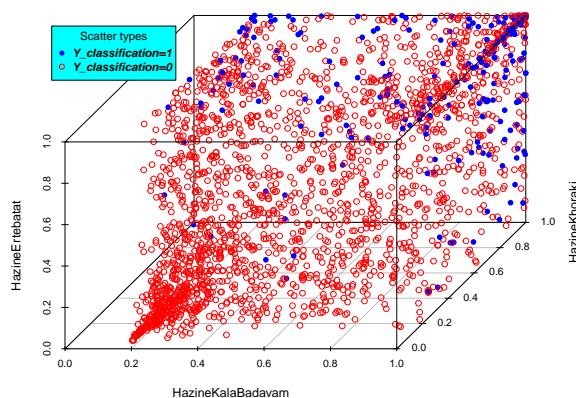
بر اساس خروجی این پسته، مهم ترین ویژگی های ما به شرح مقابل هستند.

```
1 ['lebasshoyi',
2 'panke',
3 'mashin',
4 'HazineKalaBadavam',
5 'microfer',
6 'Y_classification_mean_by_SarparstFaaliat',
7 'zarfshoyi',
8 'HazineErtebatat',
9 'Y_classification_mean_by_SarparstZanashoyi',
10 'coolerabisabet',
11 'radiozabt',
12 'freezer',
```

من در این بخش تلاش کرده ام که نموداری سه بعدی<sup>۵۹</sup> از سه ویژگی هم بسته با متغیر پاسخ رسم کنم. اما در تلاش اولیه هیچ نظم خاصی میان داده های رده توفيق و بقیه داده ها پیدا نکرد. به ناچار از تبدیلی که در رسم نمودارهای جعبه ای پهلو به پهلو استفاده کرده بودم، کمک گرفتم. این تبدیل که ساخته و پرداخته ذهن خودم هست انتقال و تجانسی از تابع sigmoid می باشد که پراکندگی<sup>۵۷</sup> داده هایی که در بازه ۰ و ۱ قرار دارند را از هم باز می کند و به پراکندگی داده ها در این بازه صورت یکنواخت تری می دهد. من صرفا جهت یاد آوری معادله ریاضی تابع sigmoid را در ادامه قرار خواهم داد.

$$\text{sigmoid}(x) = \frac{1}{1 - e^{-x}}$$

من برای توضیح بیشتر از یک رمزینه پاسخ سریع استفاده کرده ام. شما نیز می توانید با گرفتن دوربین تلفن همراهتان به سمت این تصویر وارد صفحه ای شوید که به صورت پویا عملکرد این تبدیل را نشان خواهد داد.



شکل ۴۳: نمودار پراکنش سه بعدی

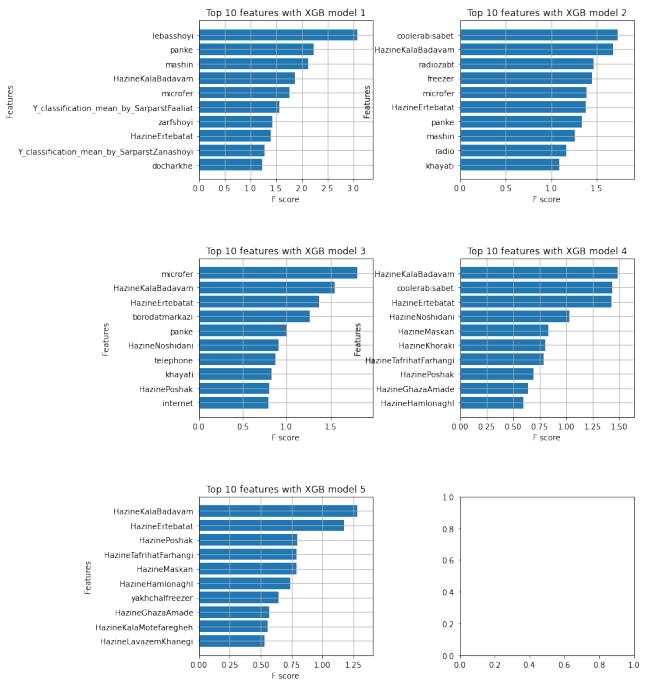
شکل ۴۳:

واضح است که در این نمودار عمدتاً خانوار ای دهک برتر اقتصادی در مراتب خارجی این نمودار پراکنش<sup>۵۸</sup> قرار دارند. این

<sup>56</sup>Three-Dimensional

<sup>57</sup>Dispersion

<sup>58</sup>Scatter Diagram



شکل ۴۴: اهمیت ویژگی‌ها با اعمال الگوریتم گرادیان بوستینگ در اجراهای مختلف

```

13 'radio',
14 'borodatmarkazi',
15 'HazineNoshidani',
16 'telephone',
17 'khayati',
18 'HazinePoshak',
19 'HazineMaskan',
20 'HazineKhoraki',
21 'HazineTafrihatFarhangi',
22 'HazineGhazaAmade',
23 'HazineHamlonagh',
24 'yakhchalfreezer',
25 'HazineKalaMotecareghch'

```

ممکن است این سؤال پیش بیايد که چرا کاهش <sup>بعد</sup> را به کمک رگرسیون لجستیک<sup>۶۴</sup>، تجزیه و تحلیل مؤلفه اصلی<sup>۶۵</sup> و یا الگوریتم هایی مثل درخت تصمیم انجام نداده‌ام؟ من همه حالات را بررسی کردم و نتیجه نهایی به کمک این انتخاب ویژگی با توجه به آسان کردن شرایط جمع‌آوری اطلاعات قابل قبول تر از سایر روش‌ها بود.

<sup>64</sup>Logistic Regression

<sup>65</sup>Principal Component Analysis

## ۳ رگرسیون لجستیک

```
34 Data.valid<-X_standard_scaler.valid[,impv]
35 Data.test<-X_standard_scaler.test[,impv]
```

با اجرای دستورات پیشین، دست یافته هایی که در بخش انتخاب ویژگی به آنها رسیده بودم را استفاده می کنم و همه پیشگوها را وارد مدل نمی کنم.

```
1 library(ROCR)
2 library(grid)
3 library(broom)
4 library(caret)
5 library(tidyr)
6 library(dplyr)
7 library(scales)
8 library(ggplot2)
9 library(ggthemr)
10 library(ggthemes)
11 library(gridExtra)
12 library(data.table)
13
14 model_glm <- glm( Y_classification ~ . ,
  data = Data.train, family = binomial(
  logit) )
15 summary_glm <- summary(model_glm)
16
17 list( summary_glm$coefficient,
18       round( 1 - ( summary_glm$deviance /
  summary_glm>null.deviance ), 2 ) )
```

در ادامه گزارشی از مدل را بینیم و ادامه بدھیم تا اهمیت هر پیشگو با کمک پی-مقدار<sup>۶۸</sup> ها مشخص شوند.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-۰.۴۷۵۳	۱۲۹۸۸۴	-۴.۵۳	۲.۷۴ × ۱۰⁻۵ ***
lebabshoyi	-۰.۶۸۷۰	۰.۶۸۸۹	-۱.۰۵	۰.۲۷۱۵
panke	-۰.۴۸۵۴	۰.۳۸۰۳	-۱.۷۶	۰.۱۸۶۹
mashin	۰.۱۱۷۹	۰.۶۵۶۷	۰.۱۷۳	۰.۹۰۰۱۸۶۷ ***
HazineKalaBadavam	۰.۷۵۴۳	۰.۲۳۵۱	۳.۶۱	۰.۰۰۱۳۷۷ **
microfer	۰.۶۷۶۹	۰.۳۷۸۴	۲.۷۶	۰.۰۰۷۴۸۰ **
Y_classification_mean_by_SarparstFaaliat	۰.۷۹۰۱	۰.۶۴۴۴	۱.۱۹	۰.۲۷۰۷
zarfshoyi	۰.۱۱۴۸	۰.۳۷۴۴	۰.۳۷	۰.۷۰۷۰
HazineErtebatat	۰.۰۱۹۱	۰.۱۷۴۱	۰.۹۷۸	۰.۹۰۰۲۹۰ **
Y_classification_mean_by_SarparstZanashoyi	-۰.۱۰۲۰	۰.۱۷۴۵	-۱.۰۵	۰.۲۷۰۸
coolerabisabet	۰.۰۸۴۶	۰.۲۷۲۶	۰.۲۸	۰.۷۰۵۷
radiozabt	۰.۰۷۷۹	۰.۳۷۸۸	۰.۱۶	۰.۷۴۴۵
freezer	۰.۰۷۵۲	۰.۵۹۵۳	۰.۷۹	۰.۴۷۷۴۸۹
radio	۰.۱۰۴۴	۰.۱۹۲۱	۰.۵۱	۰.۷۳۷۵
borodatmarkazi	-۰.۰۳۰۱	۰.۳۹۴۲	-۰.۸۹	۰.۳۷۳۲۷۱
HazineNoshidani	۰.۰۹۸۸	۰.۳۷۷۶	۰.۲۴	۰.۷۰۲۷۱ **
telephone	۰.۱۰۵۵	۰.۳۵۲۹	۰.۲۷	۰.۷۰۰۳۴ **
khayati	۰.۰۳۰۱	۰.۳۷۳۸	۰.۱۸۸	۰.۷۴۷۴۷ *
HazinePoshak	۰.۰۲۵۱	۰.۰۹۳۲	۲.۱۰	۰.۰۷۸۲۷۴
HazineMaskan	۰.۰۸۷۹	۰.۲۸۸۲	۰.۲۱	۰.۷۰۰۲۹۹ **
HazineKhoraki	۰.۰۶۶۲	۰.۱۶۱۸	۰.۲۸۰	۰.۷۰۰۳۷۶ **
HazineTafrihatFarhangi	-۰.۰۰۷۸	۰.۰۸۷۸	-۰.۰۷	۰.۷۰۸۱۶۹
HazineGhazaAmade	۰.۰۲۴۱	۰.۰۹۹۸	۰.۲۱	۰.۷۰۶۳۱۴ *
HazineHamlonaghl	۰.۰۱۹۱	۰.۱۰۰۴	۰.۱۰	۰.۷۱۸۱۷۱
yakhchalfreezer	۰.۰۶۷۶	۰.۰۸۳۷	۰.۰۸۱	۰.۷۰۲۳۱۰
HazineKalaMotefaregheh	-۰.۰۱۲۴	۰.۰۹۳۷	-۰.۱۱۳	۰.۷۰۲۰۳۹
prediction	-۰.۴۳۶۷	۰.۳۸۷۲	-۱.۱۱۲	۰.۷۰۸۴۹۶

```
1 Data.train$prediction <- predict( model_glm,
  newdata = Data.train, type = "response"
```

<sup>۶۸</sup>P-Value

اکنون آماده هستیم تا اولین مدل رده بندی را روی داده های مجموعه آموزشی اعمال کنیم. همان طور که نام فصل گویاست از مدل رگرسیون لجستیک برای پیشگویی استفاده می کنیم. من خیلی راحت می توانستم با چند دستور ساده مدل رگرسیون لجستیک را اجرا کنم. اما از این کار اجتناب می کنم. تلاش می کنم تا مقدار بُرینشی<sup>۶۶</sup> بهینه را نیز به دست آورم برای این منظور مجبور به استفاده از کتابخانه های زیاد به همراه ۳ تابع هستم که آنها را تعریف خواهم کرد.

برای اجتناب از تغییر داده های افزایش شده متغیر جدید ایجاد می کنم و مدل را روی متغیرهای جدید آموزش می دهم.

```
1 importants<-c('lebabshoyi',
2   'panke',
3   'mashin',
4   'HazineKalaBadavam',
5   'microfer',
6   'Y_classification_mean_by_SarparstFaaliat',
7   '',
8   'zarfshoyi',
9   'HazineErtebatat',
10  'Y_classification_mean_by_SarparstZanashoyi',
11  'coolerabisabet',
12  'radiozabt',
13  'freezer',
14  'radio',
15  'borodatmarkazi',
16  'HazineNoshidani',
17  'telephone',
18  'khayati',
19  'HazinePoshak',
20  'HazineMaskan',
21  'HazineKhoraki',
22  'HazineTafrihatFarhangi',
23  'HazineGhazaAmade',
24  'HazineHamlonaghl',
25  'yakhchalfreezer',
26  'HazineKalaMotefaregheh')
27 impv<-c()
28 for(i in 1:length(importants)){
29   impv[i]<-which(colnames(X_standard_
30   scaler.train)==importants[i])
31 }
32 impv[length(impv)+1]<-62
33 Data.train<-X_standard_scaler.train[,impv]
```

<sup>66</sup>Cutoff Value

<sup>67</sup>Library

```

6   cm_train <- confusionMatrix( as.factor(
7     as.numeric( train[[predict]] > c )), as.
8     factor(as.numeric(train[[actual]])) )
9   cm_test  <- confusionMatrix( as.factor(
10    as.numeric( test[[predict]] > c )), as.
11    factor(as.numeric(test[[actual]])) )
12   dt <- data.table( cutoff = c,
13                     train  = cm_train$overall[["Accuracy"]],
14                     test   = cm_test$overall[["Accuracy"]])
15   return(dt)
16 })%>% rbindlist()
17 # visualize the accuracy of the train and
18 # test set for different cutoff value
19 # accuracy in percentage.
20 accuracy_long <- gather(accuracy, "data",
21   "accuracy", -1)
22 plot <- ggplot(accuracy_long, aes(cutoff,
23   accuracy, group = data, color = data))
24 geom_line(size = 1) + geom_point(size =
25   3) +
26   scale_y_continuous(label = percent) +
27   ggttitle("Train/Test Accuracy for
28   Different Cutoff") + scale_x_continuous(
29   breaks = cutoff) + theme(axis.text.x =
30   element_text(angle = 90, hjust=1, size=7))
31
32 return( list( data = accuracy, plot = plot
33   ) )
34 }
35 #Using the AccuracyCutoffInfo function:
36 accuracy_info <- AccuracyCutoffInfo(Data.
37   train, Data.valid, "prediction", "Y_
38   classification")
39 ggthemr("light")
40 accuracy_info$plot

```

حال یک تابع تعریف می‌کنم که دید تصویری خوبی نسبت به ماتریس درهم‌ریختگی<sup>۶۹</sup> به ازای مقادیر مختلف مقدار برینشی بددهد. این تابع را ConfusionMatrixInfo خواهی نامید.

```

1 ConfusionMatrixInfo <- function( data,
2   predict, actual, cutoff )
3 {
4   # extract the column ;
5   # relevel making 1 appears on the more
6   # commonly seen position in
7   # a two by two confusion matrix
8   predict <- data[[predict]]

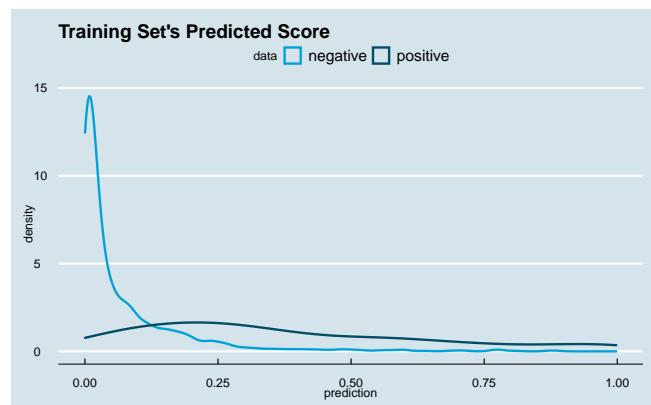
```

<sup>69</sup>Confusion Matrix

```

9   )
10  Data.valid$prediction <- predict( model_glm
11    , newdata = Data.valid , type = "
12      response" )
13
14  ggplot( Data.train, aes( prediction, color =
15    as.factor(Y_classification) ) ) +
16  geom_density( size = 1 ) +
17  ggttitle( "Training Set's Predicted Score" )
18  +
19  scale_color_economist( name = "data", labels
20    = c( "negative", "positive" ) ) +
21  theme_economist()

```



شکل ۴۵: توزیع رده توفیق و عدم توفیق در پیشگویی انجام شده توسط رگرسیون لجستیک

شکل ۴۵:

این نمودار از این حیث برای من اهمیت دارد که به من کمک کند تا دید خوبی نسبت به توزیع رده توفیق و رده عدم توفیق به دست آورم. این دید به من کمک می‌کند تا مقدار بهینه را برای مقدار بُرینشی انتخاب کنم. در دیدگاه اول به نظر می‌رسد عدد ۲۵٪ تواند انتخاب مناسبی باشد. در ادامه این انتخاب را ارزیابی خواهیم کرد.

اکنون قصد دارم یک تابع تعریف کنم که دقیق تابع را به ازای مقادیر مختلف بُرینشی روی مجموعه‌های آموزشی و اعتبار سنجی گزارش کند. سپس آن را اجرا خواهیم کرد و خروجی را که یک نمودار است در ادامه قرار خواهیم داد.

```

1 AccuracyCutoffInfo <- function( train, test,
2   predict, actual ){
3   # change the cutoff value's range as you
4   # please
5   cutoff <- seq( .3, .9, by = .01 )
6   accuracy <- lapply( cutoff, function(c){
7     # use the confusionMatrix from the caret
8     # package

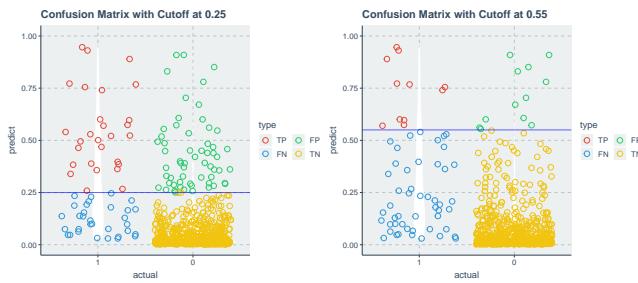
```

مقدار مختلف برای مقدار بُرینشی را به کمک تابع ConfusionMatrixInfo رسم خواهیم کرد.

```

1 cm_info <- ConfusionMatrixInfo( data = Data.
2   valid, predict = "prediction", actual =
3   "Y_classification", cutoff = .25 )
4 cm_info0 <- ConfusionMatrixInfo( data = Data.
5   valid, predict = "prediction", actual =
6   "Y_classification", cutoff = .55 )
7
8 ggarrange(cm_info$plot , cm_info0$plot)

```



شکل ۴۷: تصویری سازی ماتریس درهم‌ریختگی به ازای مقادیر بُرینشی ۰۵۵ و ۰۲۵

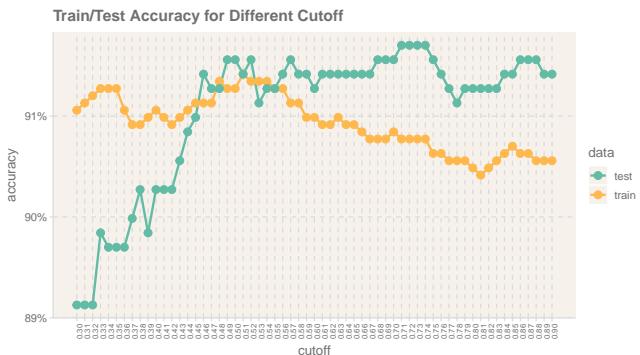
حال می‌توان به وضوح تأثیر انتخاب مقدار بُرینشی را بر دقت نهایی مشاهده کرد. اکنون می‌خواهیم با استفاده از داده‌های مجموعه آموزشی و اعتبار سنجی بهترین مقدار بُرینشی را چنان انتخاب کنم که بهترین عملکرد را داشته باشم. دقت بفرمایید که اگر بهترین مقدار بُرینشی را بر اساس تصویر ۴۶ به ازای مجموعه اعتبار سنجی انتخاب کنم، عملًا فقط روی این مجموعه دچار بیش برآش شده‌ام. لذا از همه اطلاعات به جز مجموعه تست برای این منظور استفاده می‌کنم.

برای این منظور از منحنی ROC بهره‌مند خواهیم شد لذا در این قسمت هم احتیاج به تعریف یک تابع جدید دارم.

```

1 ROCInfo <- function( data, predict, actual,
2   cost.fp, cost.fn )
3 {
4   # calculate the values using the ROCR
5   # library
6   # true positive, false positive
7   pred <- prediction( data[[predict]], data
8     [[actual]] )
9   perf <- performance( pred, "tpr", "fpr" )
10  roc_dt <- data.frame( fpr = perf@x.values
11    [[1]], tpr = perf@y.values[[1]] )
12  # cost with the specified false positive
13  # and false negative cost
14  # false positive rate * number of negative
15  # instances * false positive cost +
16  # true negative rate * number of positive
17  # instances * false negative cost
18  # calculate the cost
19  cost <- cost.fp * (roc_dt$fpr * nrow(data) -
20    sum(roc_dt$actual == 0)) + cost.fn *
21    (roc_dt$tpr * nrow(data) - sum(roc_dt$actual ==
22    1))
23 }

```



شکل ۴۶: نمودار دقت مدل به ازای مقادیر مختلف بُرینشی روی مجموعه‌های آموزشی و اعتبار سنجی (در اینجا منظور از test تست روی مجموعه اعتبار سنجی است)

```

7   actual <- relevel( as.factor( data[[actual]] ), "1" )
8   result <- data.table( actual = actual,
9     predict = predict )
# calculating each pred falls into which
# category for the confusion matrix
10  result[, type := ifelse( predict >=
11    cutoff & actual == 1, "TP",
12      ifelse( predict
13        >= cutoff & actual == 0, "FP",
14          ifelse(
15            predict < cutoff & actual == 1, "FN",
16              "TN" ) ) ] %>% as.factor() ]
# jittering : can spread the points along
# the x axis
17  plot <- ggplot( result, aes( actual,
18    predict, color = type ) ) +
19    geom_violin( fill = "white", color = NA
20    ) +
21    geom_jitter( shape=1, size = 3 ) +
22    geom_hline( yintercept = cutoff, color =
23      "blue", alpha = 0.6 ) +
24    scale_y_continuous( limits = c( 0, 1 ) ) +
25    scale_color_discrete( breaks = c( "TP",
26      "FN", "FP", "TN" ) ) + # ordering of the
27    legend
28  guides( col = guide_legend( nrow = 2 ) )
29  + # adjust the legend to have two rows
30  ggttitle( sprintf( "Confusion Matrix with
31    Cutoff at %.2f", cutoff ) )
32  return( list( data = result, plot = plot )
33  )
34 }

```

سپس تصویر ماتریس درهم‌ریختگی را به ازای دو

```

36 cost_plot <- ggplot( cost_dt, aes( cutoff,
37   cost ) ) +
38     geom_line( color = "blue", alpha =
39     0.5 ) +
40     geom_point( color = col_by_cost,
41     size = 4, alpha = 0.5 ) +
42     ggttitle( "Cost" ) +
43     scale_y_continuous( labels = comma
44   ) +
45     geom_vline( xintercept = best_
46     cutoff, alpha = 0.8, linetype = "dashed"
47   , color = "steelblue4" )
48 # the main title for the two arranged plot
49 sub_title <- sprintf( "Cutoff at %.2f -
50   Total Cost = %d, AUC = %.3f",
51   best_cutoff, best_cost, auc )
52 # arranged into a side by side plot
53 plot <- arrangeGrob( roc_plot, cost_plot,
54   ncol = 2,
55     top = textGrob( sub_title, gp =
56       gpar( fontsize = 16, fontface = "bold"
57     ) )
58 return( list( plot =
59   cutoff = best_cutoff,
60   totalcost = best_cost,
61   auc = auc,
62   sensitivity = best_tpr,
63   specificity = 1 - best_fpr ) )
64 }

```

یک نکته مهم و اساسی این است که آیا وزن اشتباهات ما در ردهبندی با هم مساوی است؟ (نرخ بد ردهبندی) یعنی اگر یارانه دهکی غیر از ۱۰ را به اشتباه قطع کنیم، با حالتی که یارانه خانوارهای دهک دهمی را به اشتباه واریز کنیم، هم هزینه هستند؟ قطعاً خیر. با توجه به سامانه اینترنتی که دولت برای ثبت اعتراضات قرار داده است، هزینه هر ۰ که به اشتباه ۱ ردهبندی شود، برابر است با هزینه کارمند متخصصی که به پرونده رسیدگی می‌کند. در مقابل، حالتی که مایک خانوار دهک دهمی را به اشتباه ۰ ردهبندی کنیم هر ماه بایستی برای او یارانه واریز کنیم. من در این مسئله فرض می‌کنم هزینه هر منفی اشتباه، (FN) ۵ برابر هر مثبت اشتباه (FP) باشد. با توجه به این فرض به بهینه‌سازی مقدار برینشی خواهیم پرداخت

```

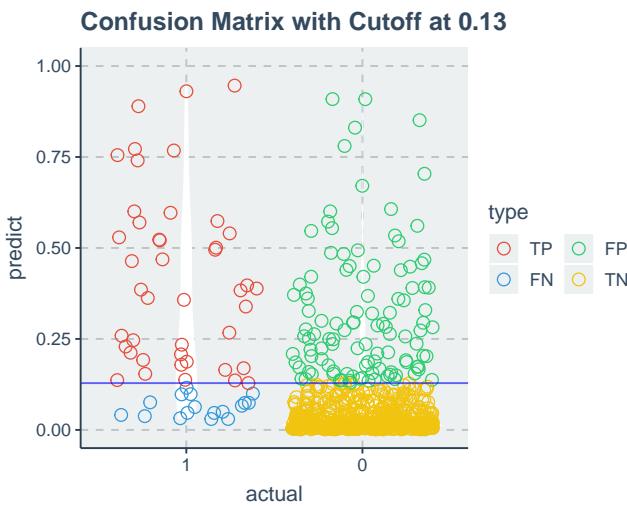
1 cost_fp <- 50
2 cost_fn <- 250
3 #As I explained in the project, I apply the
4   asymmetric cost of misclassification as
5   above.
6
7 roc_info <- ROCInfo( data = cm_info$data,
8   predict = "predict",
9

```

```

10  # false negative rate * number of positive
11  instances * false negative cost
12  cost <- perf@x.values[[1]] * cost.fp * sum
13  ( data[[actual]] == 0 ) +
14  ( 1 - perf@y.values[[1]] ) * cost.fn *
15  sum( data[[actual]] == 1 )
16  cost_dt <- data.frame( cutoff =
17  pred@cutoffs[[1]], cost = cost )
18  # optimal cutoff value, and the
19  corresponding true positive and false
20  positive rate
21  best_index <- which.min(cost)
22  best_cost <- cost_dt[ best_index, "cost"
23  ]
24  best_tpr <- roc_dt[ best_index, "tpr" ]
25  best_fpr <- roc_dt[ best_index, "fpr" ]
26  best_cutoff <- pred@cutoffs[[1]][ best_
27  index ]
28  # area under the curve
29  auc <- performance( pred, "auc" )@y.values
30  [[1]]
31  # normalize the cost to assign colors to 1
32  normalize <- function(v) ( v - min(v) ) /
33  diff( range(v) )
34  # create color from a palette to assign to
35  the 100 generated threshold between 0 ~
36  1
37  # then normalize each cost and assign
38  colors to it, the higher the blacker
39  # don't times it by 100, there will be 0
40  in the vector
41  col_ramp <- colorRampPalette( c( "green",
42  "orange", "red", "black" ) )(100)
43  col_by_cost <- col_ramp[ ceiling(
44  normalize(cost) * 99 ) + 1 ]
45  roc_plot <- ggplot( roc_dt, aes( fpr, tpr
46  ) +
47    geom_line( color = rgb( 0, 0, 1,
48    alpha = 0.3 ) ) +
49    geom_point( color = col_by_cost,
50    size = 4, alpha = 0.2 ) +
51    geom_segment( aes( x = 0, y = 0,
52    xend = 1, yend = 1 ), alpha = 0.8, color
53    = "royalblue" ) +
54    labs( title = "ROC", x = "False
55    Positive Rate", y = "True Positive Rate"
56  ) +
57    geom_hline( yintercept = best_tpr,
58    alpha = 0.8, linetype = "dashed", color
59    = "steelblue4" ) +
60    geom_vline( xintercept = best_fpr,
61    alpha = 0.8, linetype = "dashed", color
62    = "steelblue4" )

```



```

6           actual = "actual", cost
7             .fp = cost_fp, cost.fn = cost_fn )
8
9 grid.draw(roc_info$plot)
10 print(roc_info$cutoff)
11 print((roc_info$sensitivity+roc_info$specificity)/2)
12 print(roc_info$auc)
13 #-----
14 [1] 0.1286086
15 [1] 0.7722613
16 [1] 0.8575117

```

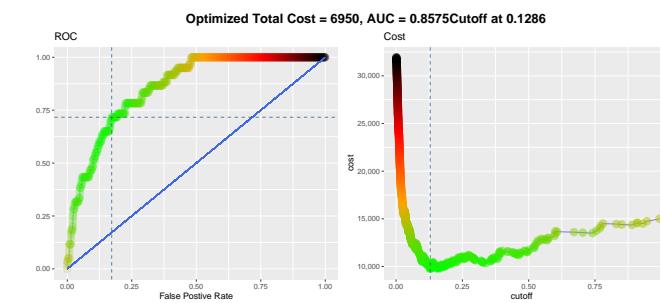
شکل ۴۹: تصویر ماتریس درهم‌یختگی به ازای مقدار بُرینشی بهینه شده بر اساس هزینه بد رده‌بندی

```

4 Data.test$prediction <- predict( model_glm,
      newdata = Data.test , type = "response"
    )

5
6 #apply optimized cut-off value
7 for(i in 1:dim(Data.train)[1]){
8   if(Data.train$prediction[i]>=roc_info$cutoff){
9     Data.train$prediction[i]<-1
10 } else{
11   Data.train$prediction[i]<-0
12 }
13 }
14 for(i in 1:dim(Data.valid)[1]){
15   if(Data.valid$prediction[i]>=roc_info$cutoff){
16     Data.valid$prediction[i]<-1
17   } else{
18     Data.valid$prediction[i]<-0
19   }
20 }
21 for(i in 1:dim(Data.test)[1]){
22   if(Data.test$prediction[i]>=roc_info$cutoff){
23     Data.test$prediction[i]<-1
24   } else{
25     Data.test$prediction[i]<-0
26   }
27 }
28 #-----#
30 > confusionMatrix(as.factor(Data.train$prediction) , as.factor(Data.train$Y_

```



شکل ۴۸: نمودارهای بهینه‌سازی مقدار بُرینشی

شکل ۴۸:

پس از این مرحله متر balanced accuracy که میانگین دو معیار مهم ارزیابی عملکرد مدل به نام‌های حساسیت<sup>۷۰</sup> و تشخیص<sup>۷۱</sup> است، از حدود ۶۲٪ به ۷۷٪ افزایش پیدا کرد. این کار نمونه بارز بهینه‌سازی فرآپارامتر هاست.

به عبارت دیگر نه تنها با توجه به هزینه بد رده‌بندی<sup>۷۲</sup>، اقدام به بهینه‌سازی مقدار بُرینشی کردم، بلکه میانگین دو معیار ارزیابی عملکرد پژوهه با بیش از ۲۴ درصد افزایش دادم. حال گزارش عملکرد الگوریتم لجستیک نهایی را روی هر سه مجموعه آموزشی، اعتبار سنجی و تست را ارائه خواهیم کرد. فقط صرفاً جهت یادآوری تکرار می‌کنم که هیچ استفاده‌ای در فرایند انتخاب بهترین مدل از دقت گزارش شده روی مجموعه تست نخواهیم کرد.

```

1
2 Data.train$prediction <- predict( model_glm,
      newdata = Data.train , type = "response"
    )
3 Data.valid$prediction <- predict( model_glm
      , newdata = Data.valid , type = "
        response" )

```

<sup>70</sup>Sensitivity

<sup>71</sup>Specificity

<sup>72</sup>Misclassification

```

81      Balanced Accuracy : 0.7723
82
83      'Positive' Class : 0
84
85 > confusionMatrix(as.factor(Data.test$prediction), as.factor(Data.test$Y_classification))
86 Confusion Matrix and Statistics
87
88      Reference
89 Prediction 0 1
90      0 168 9
91      1 36 21
92
93      Accuracy : 0.8077
94      95% CI : (0.7513, 0.8561)
95 No Information Rate : 0.8718
96 P-Value [Acc > NIR] : 0.9979475
97
98      Kappa : 0.3783
99
100 Mcnemar's Test P-Value : 0.0001063
101
102      Sensitivity : 0.8235
103      Specificity : 0.7000
104      Pos Pred Value : 0.9492
105      Neg Pred Value : 0.3684
106      Prevalence : 0.8718
107      Detection Rate : 0.7179
108      Detection Prevalence : 0.7564
109      Balanced Accuracy : 0.7618
110
111      'Positive' Class : 0

```

به عبارت دیگر دقت balanced accuracy مدل رگرسیون لجستیک روی مجموعه اعتبار سنجی برابر ۰.۷۷۲۳٪ می باشد. لازم به ذکر است در گزارش فوق می توان به جزئیات بیشتری دست یافت.

```

classification))
31 Confusion Matrix and Statistics
32
33      Reference
34 Prediction 0 1
35      0 1057 30
36      1 198 113
37
38      Accuracy : 0.8369
39      95% CI : (0.8165, 0.8559)
40 No Information Rate : 0.8977
41 P-Value [Acc > NIR] : 1
42
43      Kappa : 0.4159
44
45 Mcnemar's Test P-Value : <2e-16
46
47      Sensitivity : 0.8422
48      Specificity : 0.7902
49      Pos Pred Value : 0.9724
50      Neg Pred Value : 0.3633
51      Prevalence : 0.8977
52      Detection Rate : 0.7561
53      Detection Prevalence : 0.7775
54      Balanced Accuracy : 0.8162
55
56      'Positive' Class : 0
57 > confusionMatrix(as.factor(Data.valid$prediction), as.factor(Data.valid$Y_classification))
58 Confusion Matrix and Statistics
59
60      Reference
61 Prediction 0 1
62      0 529 17
63      1 110 43
64
65      Accuracy : 0.8183
66      95% CI : (0.7877, 0.8462)
67 No Information Rate : 0.9142
68 P-Value [Acc > NIR] : 1
69
70      Kappa : 0.3199
71
72 Mcnemar's Test P-Value : 3.25e-16
73
74      Sensitivity : 0.8279
75      Specificity : 0.7167
76      Pos Pred Value : 0.9689
77      Neg Pred Value : 0.2810
78      Prevalence : 0.9142
79      Detection Rate : 0.7568
80      Detection Prevalence : 0.7811

```

۴ مدل درخت تصمیم

الگوريتم درخت تصميم<sup>۷۳</sup> دارای فرا پaramتر هاي است که احتياج به تنظيم<sup>۷۴</sup> فرا پaramترها يا بهينه سازي فرا پaramتر<sup>۷۵</sup> دارند. يك راه ساده جستجوی شبکه اي<sup>۷۶</sup> است. در اين روش با جايگزين کردن بخشی از فرا پaramترها تلاش می کنيم فرا پaramترهاي که بيشترین دقت را روی مجموعه اعتبار سنجي کسب کرده اند را پيدا کنيم. سپس مدل را به عنوان مدل نهايی در اين بخش (ونه در كل پروژه) معرفی کنيم.

```
27 #compute the balanced accuracy
28 bla_val_dt<-c()
29 for(i in 1:num_models){
30   a<-as.factor(predict(dt_model_grid_
31   search[[i]], newdata =Data.valid, type="class"))
32   b<-as.factor(Data.valid$Y_classification)
33   bla_val_dt[i]<-as.numeric(
34     confusionMatrix(a,b)$byClass)[11]
35 }
36 #identify the best model
37 best_dt_model<-dt_model_grid_search[[which.
38   max(bla_val_dt)]]
```

تا این مرحله بهترین مدل را روی داده‌های اعتبارسنجی پیدا کردیم. ابتدا یک گزارش می‌گیریم تا فرا پارامترهای منتخب مشخص شوند.

```
1 > best_dt_model$control
2 $minsplit
3 [1] 20
4 $minbucket
5 [1] 7
6 $cp
7 [1] 0.005524272
8 $maxcompete
9 [1] 4
10 $maxsurrogate
11 [1] 5
12 $usesurrogate
13 [1] 2
14 $surrogatestyle
15 [1] 0
16 $maxdepth
17 [1] 6
18 $xval
19 [1] 10
```

حال وقت آن رسیده است که دقت مدل نهایی این قسمت را روی مجموعه اعتبار سنجی بسنجم.

```

1 > y_pred<-predict(best_dt_model,newdata=Data
2   .train,type="class")
2 > confusionMatrix(as.factor(y_pred),as.
3   factor(Data.train$Y_classification))
3 Confusion Matrix and Statistics
4
5             Reference
6 Prediction      0      1
7           0 1224    63
8           1   31    80
9

```

```

1 #DT
2 Data.train<-X_standard_scaler.train[,impv]
3 Data.valid<-X_standard_scaler.valid[,impv]
4 Data.test<-X_standard_scaler.test[,impv]
5
6 i<-26
7 Data.train[,i]<-as.factor(Data.train[,i])
8 Data.valid[,i]<-as.factor(Data.valid[,i])
9 Data.test[,i]<-as.factor(Data.test[,i])
10
11 #Define hyperparameters for grid search
12 parms<-c("information","gini")
13 maxdepth<-seq(1,30,1)
14 cp_float<-seq(0,45,1) #This variable must be
   a float number. Due to the limitations
   of the expand.grid function, I will
   modify this parameter later.
15 hyperparam_grid_dt_model<-expand.grid(parms=
   parms,maxdepth=maxdepth,cp=cp_float)
16 num_models<-nrow(hyperparam_grid_dt_model)
17 dt_model_grid_search<-list()
18
19 #Apply all models
20 for(i in 1:num_models){
21   minsplit<-hyperparam_grid_dt_model$minsplit[i]
22   maxdepth<-hyperparam_grid_dt_model$maxdepth[i]
23   cp_float<-1/((sqrt(2))^(hyperparam_grid_dt_model$cp[i]))
24   dt_model_grid_search[[i]]<-rpart(formula=Y_classification~.,
   data=Data.train,
   method="class" ,control=rpart.control(
   maxdepth=maxdepth, cp =cp_float,parms=
   parms))
25 }
26

```

73 Decision Tree

## Decision Tree 74 Hyperparameter

Hyperfine  
75Tuning

## Fining <sup>76</sup>Hyperparameter Optimization

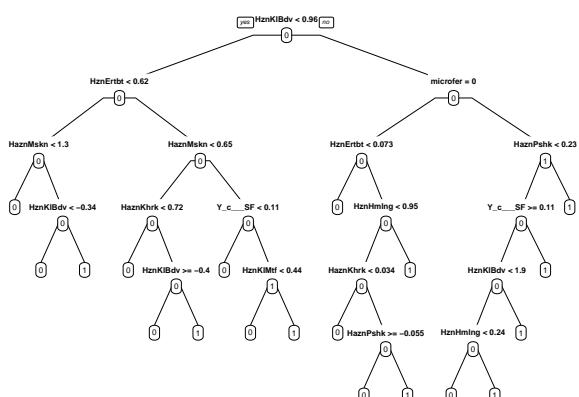
## Hyperparallel $^{77}\text{Grid Search}$

```

60 > confusionMatrix(as.factor(y_pred),as.
  factor(Data.test$Y_classification))
61 Confusion Matrix and Statistics
62
63     Reference
64 Prediction   0   1
65       0 195 18
66       1   9 12
67
68             Accuracy : 0.8846
69             95% CI : (0.8366, 0.9226)
70             No Information Rate : 0.8718
71             P-Value [Acc > NIR] : 0.3192
72
73             Kappa : 0.4081
74
75 Mcnemar's Test P-Value : 0.1237
76
77             Sensitivity : 0.9559
78             Specificity : 0.4000
79             Pos Pred Value : 0.9155
80             Neg Pred Value : 0.5714
81             Prevalence : 0.8718
82             Detection Rate : 0.8333
83             Detection Prevalence : 0.9103
84             Balanced Accuracy : 0.6779
85
86 'Positive' Class : 0

```

لذا دقت balanced accuracy ما روی مجموعه اعتبار سنجی برابر ۶۱۸۷٪ شد. اما از آنجایی که از درخت تصمیم می‌توان نمودار رسم کرد که به کمک آنها به درک درست‌تری از ویژگی‌ها بررسیم، این فصل را در این نقطه تمام نمی‌کنم و قبل از آن این نمودارها را بررسی خواهم کرد.



شکل ۵۰: نمودار نهایی درخت تصمیم

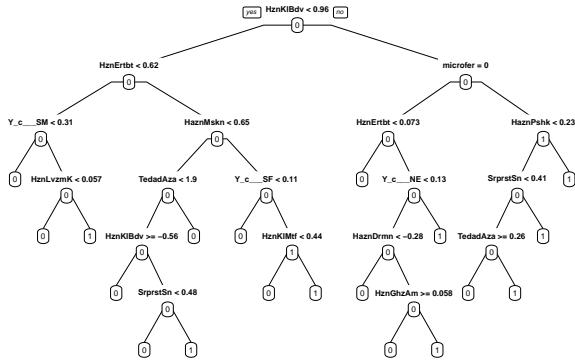
همان‌طور که در شکل ۵۰ نیز می‌توان دید، مهم‌ترین ویژگی‌ها

```

10             Accuracy : 0.9328
11             95% CI : (0.9183, 0.9453)
12             No Information Rate : 0.8977
13             P-Value [Acc > NIR] : 3.071e-06
14
15             Kappa : 0.5936
16
17 Mcnemar's Test P-Value : 0.001387
18
19             Sensitivity : 0.9753
20             Specificity : 0.5594
21             Pos Pred Value : 0.9510
22             Neg Pred Value : 0.7207
23             Prevalence : 0.8977
24             Detection Rate : 0.8755
25             Detection Prevalence : 0.9206
26             Balanced Accuracy : 0.7674
27
28 'Positive' Class : 0
29
30 > y_pred<-predict(best_dt_model,newdata=Data
  .valid,type="class")
31 > confusionMatrix(as.factor(y_pred),as.
  factor(Data.valid$Y_classification))
32 Confusion Matrix and Statistics
33
34     Reference
35 Prediction   0   1
36       0 599 42
37       1  40 18
38
39             Accuracy : 0.8827
40             95% CI : (0.8565, 0.9056)
41             No Information Rate : 0.9142
42             P-Value [Acc > NIR] : 0.9982
43
44             Kappa : 0.241
45
46 Mcnemar's Test P-Value : 0.9121
47
48             Sensitivity : 0.9374
49             Specificity : 0.3000
50             Pos Pred Value : 0.9345
51             Neg Pred Value : 0.3103
52             Prevalence : 0.9142
53             Detection Rate : 0.8569
54             Detection Prevalence : 0.9170
55             Balanced Accuracy : 0.6187
56
57 'Positive' Class : 0
58
59 > y_pred<-predict(best_dt_model,newdata=Data
  .test,type="class")

```

به شرح ذیل است.



شکل ۵۱: نمودار بهترین درخت تصمیم در صورتی که از کاهش بُعد استفاده نمی‌کردیم

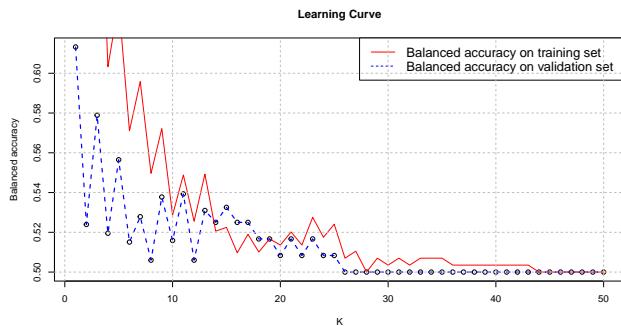
به خوبی انجام شده بود. به هر حال من مدل نهایی را همان چیزی که قبلاً اعلام کرده بودم در بخش درخت تصمیم در نظرخواهم گرفت.

1	HazineKalaBadavam
2	49.9075460
3	HazineErtebatat
4	21.7005220
5	HazineKhoraki
6	15.9740547
7	HazineMaskan
8	14.2443350
9	HazineHamlonaghl
10	12.2060157
11	HazineKalaMotefaregheh
12	9.5608606
13	HazinePoshak
14	8.3819017
15	microfer
16	8.2325922
17	zarfshoyi
18	5.4561673
19	HazineGhazaAmade
20	5.3440773
21	Y_classification_mean_by_SarparstFaaliat
22	5.1494505
23	HazineNoshidani
24	2.5714052
25	HazineTafrihatFarhangi
26	2.5522031
27	telephone
28	1.7564724
29	panke
30	1.3787546
31	radiozabt
32	1.0209799
33	Y_classification_mean_by_SarparstZanashoyi
34	0.8686019
35	borodatmarkazi
36	0.8592599
37	freezer
38	0.3589744
39	yakhchalfreezer
40	0.3589744
41	mashin
42	0.3257143

نکته مهم آن است که اگر کاهش بُعد را انجام نمی‌دادم، دقت نهایی ۳.۰ درصد بیش تر می‌شد. اما به عقیده من کاهش قابل توجه بُعد می‌توانست ارزش آن را داشته باشد که این کاهش بُعد را انجام دهم. ضمناً مدل نهایی انتخاب شده در حالتی که کاهش بُعد نداشتیم خیلی تفاوت چندانی با مدل کنونی ندارد. (با توجه به بررسی نمودارهای درخت تصمیم) با مقایسه این دو نمودار نیز می‌توان متوجه شد فرایند کاهش بعد

## ۵ مدل k نزدیک‌ترین همسایه

سومین الگوریتمی که قرار است روی داده‌ها اعمال کنم الگوریتم K نزدیک‌ترین همسایه<sup>۱۸</sup> است. چالش مهم من در این الگوریتم محاسبه K مناسب است. لذا بایستی به کمک داده‌های مجموعه اعتبار سنجی، اقدام به بهینه‌سازی فرا پارامتر نماییم.



شکل ۵۲: منحنی یادگیری

البته به شخصه ترجیح می‌دادم عددی به جز ۱ بهینه‌ترین مقدار برای K باشد. چراکه یکی از هشدارهایی که برای بیش‌برازش در مقاله ارجاع شده مطرح شده است کوچک بودن مقدار بهینه برای K است.

```

1 #knn
2 Data.train<-X_standard_scaler.train[,impv]
3 Data.valid<-X_standard_scaler.valid[,impv]
4 Data.test<-X_standard_scaler.test[,impv]
5
6 i<-26
7 Data.train[,i]<-as.factor(Data.train[,i])
8 Data.valid[,i]<-as.factor(Data.valid[,i])
9 Data.test[,i]<-as.factor(Data.test[,i])
10
11 bla_val_knn<-c()
12 for(i in 1:50){
13   y_pred = knn(train = Data.train[,1:25] ,
14                 test = Data.valid[,1:25] ,
15                 cl = Data.train$Y_classification ,
16                 k = i)
17   a<-as.factor(y_pred)
18   b<-as.factor(Data.valid$Y_
19   classification)
20   bla_val_knn[i]=as.numeric(
21     confusionMatrix(a,b)$byClass)[11]
20 }
```

اکنون `balanced accuracy` همه حالت‌های ممکن در بردار `bla_val_knn` ذخیره شده است. حال برای به دست آمدن نتایج خواهیم داشت:

```

1 > which.max(bla_val_knn)
2 [1] 1
3 > bla_val_knn[which.max(bla_val_knn)]
4 [1] 0.6132238
```

به عبارتی من با قراردادن  $K = 1$  بهترین نتیجه را از این مدل می‌گیرم. `balanced accuracy` در مدل نهایی من برابر  $61\%^{۷۸}$  است.

یکی از مهم‌ترین نمودارهایی که برای این الگوریتم همواره رسم می‌شود نمودار منحنی یادگیری<sup>۱۹</sup> است. من به کمک این نمودار نه تنها می‌توانم مقدار K بهینه را پیدا کنم. بلکه این نمودار تا حدی می‌تواند راهنمای من برای جلوگیری از بیش‌برازش<sup>۲۰</sup> باشد. [۱۱]

<sup>78</sup>K-Nearest Neighbor

<sup>79</sup>Learning Curve

<sup>80</sup>Overfitting

## ۶ مدل شبکه عصبی

```

29         rep=1,
30         threshold=0.05,
31         learningrate=LR[i],
32         startweights="NULL",
33         algorithm=algorithm[AL[i]],
34         act.fct=activation_function[
      AF[i]],
35             stepmax=2*10^5)
36             if(is.numeric(dim(nn$result.
matrix)[1])) {
37                 y_pred_nn<-compute(nn,
nn.valid[,1:25])
38                 a<-as.factor(apply(y_
pred_nn$net.result,1,which.max)-1)
39                 b<-as.factor(nn.valid$Y_
classification)
40                 bla_val_nn[i]<-as.
numeric(confusionMatrix(a,b)$byClass)
[11]
41                 HL[i]<-hidden_layer
42             }
43 }
```

از آنجایی که شبکه عصبی به راحتی دچار بیش برآش می شود من ۱۵۰ بار الگوریتم را با بردارهایی که نمایانگر لایه پنهان<sup>۸۴</sup> های مختلف و تصادفی (چه از نظر تعداد گرهها چه از نظر تعداد لایهها) هستند اجرا کردم. همچنین از دیگر فرآپارامترهای مختلف اعم از توابع فعال ساز و نرخ یادگیری<sup>۸۵</sup> های تصادفی و ... در هر حلقه استفاده کرده‌ام. بیشترین دقیقی که روی مجموعه اعتبار سنجی حاصل شد را به عنوان مدل نهایی در نظر گرفتم.

```

1 > best_i<-where.min[bla_val_nn]
2 >HL[best_i] #To display the best hiddenlayer
   vector
3 [[1]]
4 [1] 13 5
5 >activation_function[AF[best_i]] #To display
   the best Activation Function
6 [1] "logistic"
7 >algorithm[AL[best_i]] #To display the best
   Algorithm
8 [1] "sag"
9 >LR[best_i] #To display the best Learning
   Rate
10 [1] 0.4789158
11 >bla_val_nn[best_i]
12 [1] 0.6830203
```

در نتیجه، بهترین مدل شبکه عصبی balanced accuracy را مجموعه اعتبار سنجی ۶۸۳۰۲۰٪ دارد. البته با توجه به تعداد

تلاش اولیه من این بود که بتوانم یک مدل یادگیری عمیق<sup>۸۶</sup> روی داده‌ها در زبان برنامه‌نویسی R ایجاد کنم. ولی به علت عدم تسلط کافی من روی این زبان برنامه‌نویسی این مهم محقق نشد. مشکل اساسی من در اعمال مدل ساده شبکه عصبی<sup>۸۷</sup> روی داده‌ها این بود که اولاً در تعریف تابع فعال ساز<sup>۸۸</sup> محدودیت زیادی داشتم ثانیاً توان دسترسی مناسبی به الگوریتم بهینه سازی نداشت. به هر حال با توجه به نکات گفته شده ابزار مناسبی برای استفاده حداکثری از توان شبکه عصبی برای من میسر نبود.

```

1 nn.train<-Data.train
2 nn.valid<-Data.valid
3 nn.test<-Data.test
4 #normalize to (0,1)
5 for(i in 1:dim(Data.train)[2]){
6   nn.train[[i]]<-(nn.train[[i]]-min(Data.
train[[i]]))/(max(Data.train[[i]])-min(
Data.train[[i]]))
7   nn.valid[[i]]<-(nn.valid[[i]]-min(Data.
train[[i]]))/(max(Data.train[[i]])-min(
Data.train[[i]]))
8   nn.test[[i]]<-(nn.test[[i]]-min(Data.
train[[i]]))/(max(Data.train[[i]])-min(
Data.train[[i]]))
9 #Model training based on random hyper
   parameters
10 bla_val_nn<-c()
11 HL<-list()
12 LR<-c()
13 activation_function<-c('logistic','tanh')
14 algorithm<-c('rprop+', 'rprop-','sag','slr')
15 AF<-c()
16 AL<-c()
17 for(i in 1:150){
18   print(i)
19   LR[i]<-runif(1,0,1)
20   AF[i]<-floor(runif(1,1,3))
21   AL[i]<-floor(runif(1,1,5))
22   hidden_layer_num<-floor(runif(1,1,6))
23   hidden_layer<-floor(runif(hidden_layer_
num,3,26))
24   hidden_layer<-hidden_layer[order(hidden_
layer,decreasing = TRUE)]
25   nn<-neuralnet(Y_classification~. ,data=
nn.train,
26               hidden = hidden_layer,
27               linear.output = F,
28               lifesign = 'full',
```

<sup>81</sup>Deep Learning

<sup>82</sup>Neural Network

<sup>83</sup>Activation Function

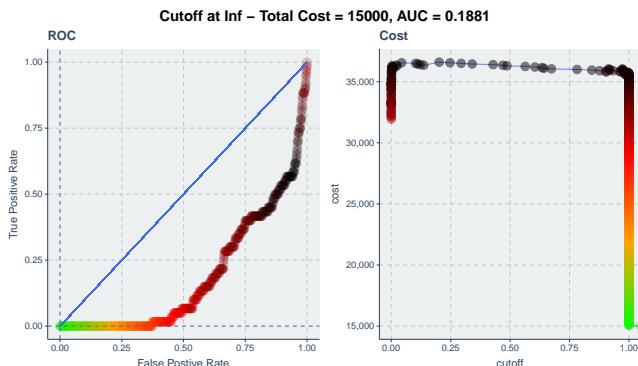
<sup>84</sup>Hidden Layer

<sup>85</sup>Learning Rate

```

2 df<-data.frame(actual=b,predict=y_pred_nn$net.result[,1],type=cm_info$data$type)
3 for(i in 1:dim(nn.valid)){
4   if(df$predict[i]>0.5){
5     if(df$actual[i]==0){
6       df$type[i]<-as.factor("TN")
7     }else{
8       df$type[i]<-as.factor("FN")
9     }
10  } else{
11    if(df$actual[i]==1){
12      df$type[i]<-as.factor("TP")
13    }else{
14      df$type[i]<-as.factor("FP")
15    }
16  }
17 }
18 #Using the ROCInfo function
19 cost_fp <- 50
20 cost_fn <- 250
21 roc_info <- ROCInfo( data = df, predict =
  "predict", actual = "actual", cost.fp =
  cost_fp, cost.fn = cost_fn )
22 #plot
23 grid.draw(roc_info$plot)

```



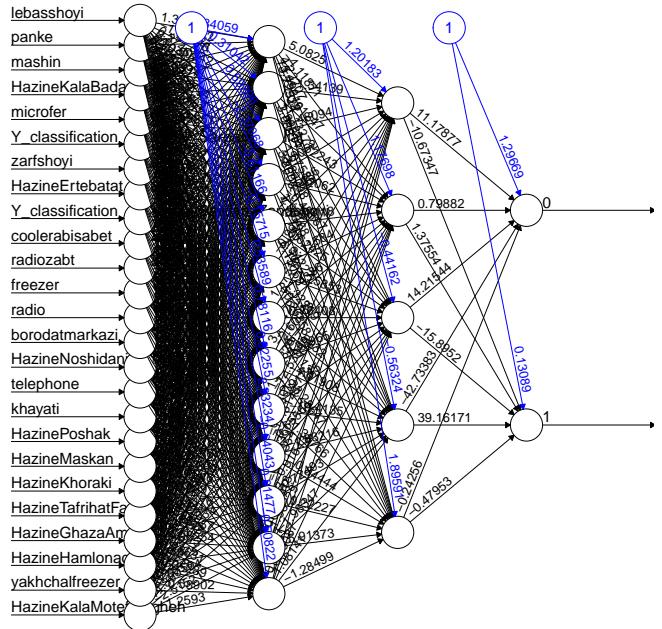
شکل ۵۴: نمودار بهینه سازی مقدار بُرینشی رده غیرتوفيق

به عبارت دیگر از شکل ۵۴ متوجه شدیم هر چقدر بیشتر مقدار بُرینشی بیشتر نزدیک ۱ بشود هزینه کمتری خواهیم پرداخت. من وضعیت balanced accuracy را بر اساس مقادیر مختلف مقدار بُرینشی بررسی خواهم کرد.

```

1 balanced_accuracy_cutoff<-c()
2 s<-0.00001
3 step<-seq(s, 1-s, by=s)
4 for(j in 1:length(step)){
5   tmp<-c()
6   for(i in 1:length(y_pred_nn$net.result[,1])){
```

کم داده‌ها نباید انتظار زیادی از این الگوریتم داشت. چرا که توان پیشگویی شبکه‌های عصبی روی داده‌های زیاد است.



اما من قصد دارم مشابه کاری که در بخش رگرسیون لجستیک انجام داده‌ام را در این بخش نیز انجام دهم.تابع هزینه را مانند همانند آنچه در بخش قبل در نظر گرفتم فرض می‌کنم. یعنی اشتباہ تشخیص دادن یک خانوار دهک دهمی را ۵ برابر هزینه‌بر از هزینه بد رده‌بندی یک خانوار دهک تا نهم فرض می‌کنم.<sup>۸۶</sup>

در نظر داشته باشید که لایه آخر مدل من در شکل ۵۳ دو گره دارد که گره اول به تعبیری احتمال صفر بودن و گره دوم احتمال یک بودن را نمایش می‌دهد. لذا اولاً جمع خروجی این دو گره همیشه ۱ می‌باشد، ثانیاً مانند خروجی شبکه عصبی آن چیزی است که احتمال بیشتری دارد. در نتیجه معادلاً می‌توان گفت، اگر احتمال صفر بودن بیشتر از نیم باشد، مدل صفر را به عنوان نتیجه رده‌بندی اعلام می‌کند. در غیر این صورت ۱ را به عنوان رگرسیون لجستیک با عنوان مقدار بُرینشی داشتیم می‌باشد.

در ابتدا تابع منحنی هزینه را با توجه به تابع هزینه مفروض و خروجی گره مربوط به رده‌بندی صفر در الگوریتم شبکه عصبی رسم می‌کنم تا مقدار بُرینشی را به بهینه‌ترین شکل انتخاب کنم. برای این منظور از توابع نوشته شده در بخش رگرسیون لجستیک بهره خواهیم برد.

```
1 #Prepare data to use the ROCInfo function
```

<sup>86</sup>Cost

```

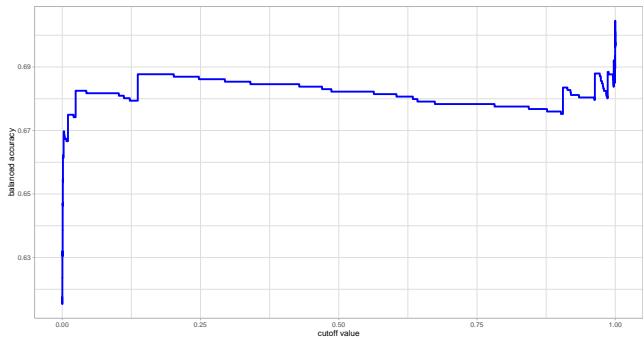
15
16 Mcnemar's Test P-Value : 9.869e-16
17
18     Sensitivity : 0.9012
19     Specificity : 0.8252
20     Pos Pred Value : 0.9784
21     Neg Pred Value : 0.4876
22     Prevalence : 0.8977
23     Detection Rate : 0.8090
24     Detection Prevalence : 0.8269
25     Balanced Accuracy : 0.8632
26
27     'Positive' Class : 0
28 #validation set
29 Confusion Matrix and Statistics
30
31     Reference
32 Prediction 0 1
33      0 549 27
34      1 90 33
35
36     Accuracy : 0.8326
37     95% CI : (0.8028, 0.8596)
38     No Information Rate : 0.9142
39     P-Value [Acc > NIR] : 1
40
41     Kappa : 0.2773
42
43 Mcnemar's Test P-Value : 9.931e-09
44
45     Sensitivity : 0.8592
46     Specificity : 0.5500
47     Pos Pred Value : 0.9531
48     Neg Pred Value : 0.2683
49     Prevalence : 0.9142
50     Detection Rate : 0.7854
51     Detection Prevalence : 0.8240
52     Balanced Accuracy : 0.7046
53
54     'Positive' Class : 0
55 #For information only
56 #I will never use this section.
57 Confusion Matrix and Statistics
58
59     Reference
60 Prediction 0 1
61      0 175 16
62      1 29 14
63
64     Accuracy : 0.8077
65     95% CI : (0.7513, 0.8561)
66     No Information Rate : 0.8718
67     P-Value [Acc > NIR] : 0.99795

```

```

7         if(y_pred_nn$net.result[,1][i]>=step[j]){
8             tmp[i]<-0
9         } else {
10             tmp[i]<-1
11         }
12     }
13 tmp<-as.factor(tmp)
14 balanced_accuracy_cutoff[j]<-as.numeric(
15     confusionMatrix(tmp,b)$byClass)[11]
16 }
17 #plot
18 ggplot(data=data.frame(balanced_accuracy_
19 cutoff), aes(x=step,y=balanced_accuracy_
20 cutoff))+geom_line(color="blue", size
21 =1.2)+xlab("cutoff value")+ylab("_
22 balanced accuracy") +theme_light()

```



شکل ۵۵: نمودار دقت - مقدار بُرینشی

ظاهراً هر چقدر آستانه مقدار بُرینشی را به یک نزدیک‌تر می‌کنیم، balanced accuracy بهبود پیدا می‌کند. بهترین دقت من برای مقدار بُرینشی ۰.۹۹۹۹۴ است که توانسته به balanced accuracy بالای ۷۰٪۴۵۷۷۵ دست یابد. اکنون ماتریس درهم‌ریختگی را برای این مقدار بُرینشی بررسی می‌کنم.

```

1 #training set
2 Confusion Matrix and Statistics
3
4     Reference
5 Prediction 0 1
6      0 1131 25
7      1 124 118
8
9     Accuracy : 0.8934
10    95% CI : (0.8761, 0.9091)
11    No Information Rate : 0.8977
12    P-Value [Acc > NIR] : 0.7196
13
14    Kappa : 0.5559

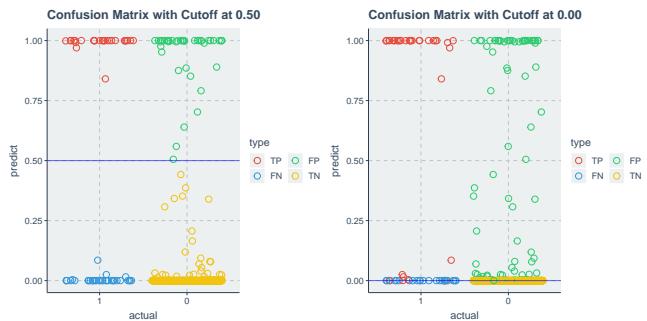
```

```

68
69           Kappa : 0.2739
70
71 McNemar's Test P-Value : 0.07364
72
73           Sensitivity : 0.8578
74           Specificity : 0.4667
75           Pos Pred Value : 0.9162
76           Neg Pred Value : 0.3256
77           Prevalence : 0.8718
78           Detection Rate : 0.7479
79           Detection Prevalence : 0.8162
80           Balanced Accuracy : 0.6623
81
82           'Positive' Class : 0

```

حال آن تصویری سازی که برای ماتریس درهم ریختگی در فصل رگرسیون لجستیک داشتم را روی نتیجه داده های مجموعه اعتبار سنجی به ازای دو مقدار بُرینشی ۰۹۹۹۹۴٪ و مقدار بُهینه ۰۹۹۹۹۴٪ را با توجه به گره مربوط به خروجی رد ۱ رسم می کنم. (لذا مقدار بُرینشی بُهینه در این حالت ۰۹۹۹۹۴٪ - ۱ خواهد بود.)



شکل ۵۶: نمودار تصویر ماتریس درهم ریختگی به ازای دو مقدار بُرینشی پیشفرض (سمت چپ) و بُهینه (سمت راست)

## انتخاب مدل نهایی

```
15 McNemar's Test P-Value : 0.0001063
16
17     Sensitivity : 0.8235
18     Specificity : 0.7000
19     Pos Pred Value : 0.9492
20     Neg Pred Value : 0.3684
21     Prevalence : 0.8718
22     Detection Rate : 0.7179
23     Detection Prevalence : 0.7564
24     Balanced Accuracy : 0.7618
```

اطلاعات فوق بهوضوح گویای دقت نهایی است و توضیحات بیشتر بی فایده است.

### فرایند انتخاب مدل

فرایند انتخاب مدل من بر اساس صفحات ۱۵۰ و ۱۵۱ کتاب Understanding Machine Learning from theory to algorithms

نوشته شای شالو شوارتز و شای بن داوید بود. [۱۲]  
با توجه به اینکه در آموزش دادن همه مدل‌ها از مجموعه اعتبارسنجی بهره بردیم، لذا اعتبار سنجی نهایی را روی مجموعه تست انجام دادم. خوشبختانه همان‌طور که مشخص است دچار بیش برآش نشدم. ضمناً این کتاب برای موقوعی که دقت ما روی مجموعه تست به شکل مشهودی کمتر از مجموعه اعتبار سنجی باشد راهکارهای خوب و مفصلی ارائه می‌دهد.

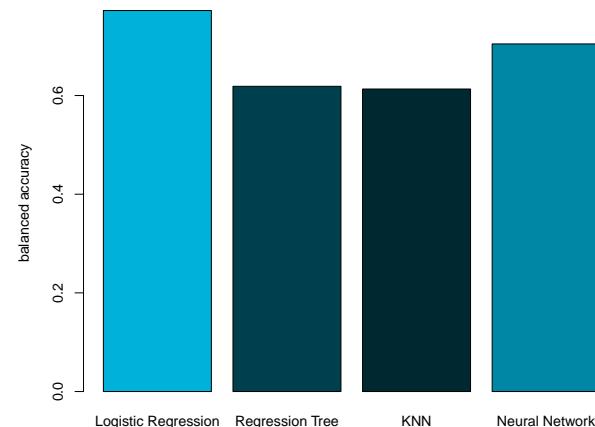
من تمام تلاش خود را کردم از همه ابزارهای ممکن در چهارچوب مرجع اصلی درس داده‌کاوی به همراه توجه به توصیه‌های استاد محترم درس بالاترین بازدهی را در این پروژه داشته باشم برای همین منظور اقداماتی مانند انتخاب ویژگی و بهینه‌سازی مقادیر بُرینشی را در خصوص الگوریتم‌هایی که امکان آن را داشتند انجام دادم. امیدوارم این اقدامات دقت نهایی من را بالاتر برده باشد.

### تقدیر و تشکر

بدین‌وسیله در وهله اول از استاد محترم درس، جناب آفای دکتر فقیهی بابت آموختن درس داده‌کاوی، تعریف پروژه و همه راهنمایی‌هایشان در طول پروژه تشکر و قدردانی می‌نمایم. باشد تا برداشت صحیحی از آنچه ایشان به ما آموختند داشته باشم. همچنین از دوستان خوبم آقایان دکتر بهراد تقی بیگلو و سهراب فریدی که من را در به سرانجام رساندن این پروژه راهنمایی کردن نهایت سپاسگزاری را دارم.

"یک تصویر به هزار کلمه می‌ارزد" این جمله برگرفته از مرجع اصلی است. لذا من از یک نمودار میله‌ای <sup>۱۳</sup> برای تصویری سازی و مقایسه میزان balanced accuracy مدل‌های منتخب در هر فصل پروژه کمک گرفتم تا در فرایند انتخاب مدل از آن بهره‌مند شوم. بهوضوح بهترین دقت مربوط به الگوریتم رگرسیون لجستیک است.

balanced accuracy of algorithms on the validation set



شکل ۵۷: مقایسه دقت مدل‌ها

البته از آنجایی که عملکرد شبکه عصبی به رگرسیون لجستیک از بقیه نزدیکتر بود من با مقایسه تصاویر ۴۸ و ۵۴ دریافت که با توجه به تابع هزینه، استفاده از تابع رگرسیون لجستیک، هزینه کمتری برای ما خواهد داشت.  
حال وقت آن رسیده است که دقت مدل منتخب نهایی را روی مجموعه تست بررسی کنیم. البته این کار را قبل از انجام داده بودیم اما هیچ استفاده‌ای از این اطلاعات نکرده بودم.

### ۱ Confusion Matrix and Statistics

```
2
3             Reference
4 Prediction   0   1
5           0 168   9
6           1  36  21
7
8                 Accuracy : 0.8077
9                 95% CI : (0.7513, 0.8561)
10                No Information Rate : 0.8718
11                P-Value [Acc > NIR] : 0.9979475
12
13                Kappa : 0.3783
14
```

<sup>۱۳</sup>Bar Chart

## مراجع

- [1] گالیت شمولی، پیترسی بروس، اینبال یاهو، نیتین آر پاتل، کنت سی داده کاوی برای تحلیل خودکار کسب و کار: مفاهیم، فنون. لیختندال و کاربردهای  $R$ .
- [2] I. M. Society, "واژه نامه ریاضی انجمن ریاضی ایران," September 2022.
- [3] S. Research and T. Center, "واژه نامه و اصطلاحات انجمن آمار," ایران, September 2022.
- [4] S. C. of Iran, "هزینه و درامد خانوارهای کل کشور," December 2020.
- [5] geetansh044., "How to normalize data in r?," December 2021.
- [6] P. Cerdà, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Machine Learning*, vol.107, no.8, pp.1477–1494, 2018.
- [7] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. New York: Springer, fourth ed. , 2002. ISBN 0-387-95457-0.
- [8] wikipedia., "Heatmap," December 2021.
- [9] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol.290, no.5500, pp.2323–2326, 2000.
- [10] P. Pudil and J. Novovičová, *Novel Methods for Feature Subset Selection with Respect to Problem Knowledge*, pp.101–116. Boston, MA: Springer US, 1998.
- [11] X. Zhu, "K-nearest-neighbor an introduction to machine learning," *Computer Sciences Department University of Wisconsin, Madison*, 2006.
- [12] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

# واژه‌نامه انگلیسی به فارسی

Distribution ..... توزیع .....

## A

F ..... Activation Function ..... تابع فعال ساز .....

Feature ..... ویژگی .....

Feature Selection ..... انتخاب ویژگی .....

First Quartile ..... چارک اول .....

Frequency ..... فراوانی .....

Function ..... تابع .....

## B

Balance ..... تعادل .....

Bar Chart ..... نمودار میله‌ای .....

Batch ..... قطعه .....

Binary ..... دو دویی .....

Boxplot ..... نمودار جعبه‌ای .....

## G

Gradient Boosting ..... گرادیان بوستینگ .....

Grid Search ..... جستجوی شبکه‌ای .....

## C

Resteهای ..... رسته‌ای .....

Compiler ..... مترجم .....

Confusion Matrix ..... ماتریس درهم‌ریختنگی .....

Correlation ..... همبستگی .....

Cost ..... هزینه .....

Cutoff Value ..... مقدار بُرینشی .....

## H

Heatmap ..... نمودار حرارتی .....

Hidden Layer ..... لایه پنهان .....

Histogram ..... بافت‌نگار .....

Hyperparameter ..... فرا پارامتر .....

Hyperparameter Optimization ..... بهینه‌سازی فرا پارامتر .....

## D

Imbalance ..... نامتعادل .....

Information ..... اطلاعات .....

Data ..... داده .....

K-Nearest Neighbor ..... K نزدیکترین همسایه .....

Data Encoding ..... کدگذاری داده‌ها .....

Decile ..... نشت داده .....

Data Leakage ..... نشت داده .....

Data Visualization ..... تصویرسازی داده‌ها .....

Decision Tree ..... درخت تصمیم .....

Deep Learning ..... یادگیری عمیق .....

Density ..... چگالی .....

Learning Curve ..... منحنی یادگیری .....

Learning Rate ..... نرخ یادگیری .....

Dimension Reduction ..... کاهش بُعد .....

Dispersion ..... پراکندگی .....

Python	پایتون	کتابخانه
		لگاریتم
		رگرسیون لجستیک
<b>Q</b>		
Quartile	چارک	
<b>R</b>		
Random	صادفی	یادگیری ماشین
Real Number	عدد حقیقی	ماتریس
Record	ثبت	میانگین
Regression	رگرسیون	میانه
Response Variable	متغیر پاسخ	بد رده بندی
		مقدار گمشده
<b>S</b>		
Scatter Diagram	نمودار پراکنش	شبکه عصبی
Sensitivity	حساسیت	نوفه
Side-By-Side Boxplot	نمودار جعبه‌ای پهلو به پهلو	توزیع نرمال
Skewness	چولگی	نرمالیده
Specificity	تشخیص	نهی
Strictly Increasing	اکیداً صعودی	کمی
Success Class	رده توفیق	
Summary	خلاصه	
Supervised	راهنماییده	
<b>T</b>		
Test Set	مجموعه آزمون	
Third Quartile	چارک سوم	افراز
Three-Dimensional	سه بعدی	چولگی مثبت
Training Set	مجموعه آموزشی	پیشگو
Transformation	تبديل	پیش‌پردازش
Tuning	تنظیم	تجزیه و تحلیل مؤلفه اصلی
		P-Value
<b>M</b>		
Machine Learning		
Matrix		
Mean		
Median		
Misclassification		
Missing Value		
<b>N</b>		
Neural Network		
Noise		
Normal Distribution		
Normalized		
Null		
Numerical		
<b>O</b>		
Overfitting		بیش‌پردازش
Oversampling		بیش‌نمونه‌گیری
<b>P</b>		
Partition		
Positive Skewness		
Predictor		
Preprocessing		
Principal Component Analysis		
P-Value		پی-مقدار

## V

مجموعه اعتبار سنجی ..... Validation Set .....

تصویری سازی ..... Visualization .....

## واژه‌نامه فارسی به انگلیسی

Categorical .....	رسته‌ای .....
Regression .....	رگرسیون .....
Logistic Regression .....	رگرسیون لجستیک .....
Three-Dimensional .....	سه بعدی .....
Neural Network .....	شبکه عصبی .....
Real Number .....	عدد حقیقی .....
Hyperparameter .....	فرا پارامتر .....
Frequency .....	فراوانی .....
Batch .....	قطعه .....
Hidden Layer .....	لایه پنهان .....
Logarithm .....	لگاریتم .....
Matrix .....	ماتریس .....
Confusion Matrix .....	ماتریس درهم‌ریختگی .....
Compiler .....	مترجم .....
Response Variable .....	متغیر پاسخ .....
Test Set .....	مجموعه آزمون .....
Training Set .....	مجموعه آموزشی .....
Validation Set .....	مجموعه اعتبار سنجی .....
Cutoff Value .....	مقدار بُرینشی .....
Missing Value .....	مقدار گم شده .....
Learning Curve .....	منحنی یادگیری .....
Median .....	میانه .....
Mean .....	میانگین .....
Imbalance .....	نامتعادل .....
Learning Rate .....	نرخ یادگیری .....
Normalized .....	نرمالیله .....
Data Leakage .....	نشت داده .....
Boxplot .....	نمودار جعبه‌ای .....
Side-By-Side Boxplot .....	نمودار جعبه‌ای پهلو به پهلو .....
Heatmap .....	نمودار حرارتی .....
Bar Chart .....	نمودار میله‌ای .....
Scatter Diagram .....	نمودار پراکنش .....
Cost .....	هزینه .....
	اطلاعات .....
	افزار .....
	انتخاب ویژگی .....
	اکیداً صعودی .....
	بافت‌نگار .....
	بد رده‌بندی .....
	بهینه‌سازی فرا پارامتر .....
	بیش‌برازش .....
	بیش‌نمونه‌گیری .....
	تابع .....
	تابع فعال ساز .....
	تبدیل .....
	تجزیه و تحلیل مؤلفه اصلی .....
	تشخیص .....
	تصادفی .....
	تصویرسازی داده‌ها .....
	تصویری سازی .....
	تعادل .....
	تنظیم .....
	تهی .....
	توزیع .....
	توزیع نرمال .....
	ثبت .....
	جستجوی شبکه‌ای .....
	حساسیت .....
	خلاصه .....
	داده .....
	درخت تصمیم .....
	دهک .....
	دودویی .....
	راهنماییده .....
	رده توفیق .....
	Success Class .....

Correlation .....	همبستگی .....
Feature .....	ویژگی .....
Python .....	پایتون .....
Dispersion .....	پراکندهگی .....
Predictor .....	پیشگو .....
Preprocessing .....	پیش‌پردازش .....
P-Value .....	پی-مقدار .....
Deep Learning .....	یادگیری عمیق .....
Machine Learning .....	یادگیری ماشین .....
K-Nearest Neighbor .....	K نزدیکترین همسایه .....
Quartile .....	چارک .....
First Quartile .....	چارک اول .....
Third Quartile .....	چارک سوم .....
Skewness .....	چولگی .....
Positive Skewness .....	چولگی مثبت .....
Density .....	چگالی .....
Dimension Reduction .....	کاهش بعد .....
Library .....	کتابخانه .....
Data Encoding .....	کدگذاری داده‌ها .....
Numerical .....	کمی .....
Gradient Boosting .....	گرادیان بوستینگ .....