Dəyişmə istiqamətinin multinomial Logit Modeli ilə proqnozlaşdırılması

written by Hikmat Abdulazizov Hikmət əbdüləzizov

Maliyyə ekonometrikası ədəbiyyatında fond birjasının gəlirlərinin müəyyən dərəcədə proqnozlaşdırıla biləcəyinə dair əhəmiyyətli dəlillər mövcuddur. Əsas məqsəd səhmlərin gəlirliliyinin ümumi səviyyəsinin şərti ortalamasını proqnozlaşdırmaq olmuşdur. Bununla birlikdə, bir neçə tədqiqat göstərdi ki, yalnız səhmlərin gəlirlilik istiqaməti proqnozlaşdırıla bilər (Kristofersen, 2006).

Gələcək gəlirliliyin işarəsinin öngörülməsi ədəbiyyatda dəyişiklik istiqaməti proqnozlaşdırılması kimi tanınır və həqiqətən treyding baxımından çox maraqlıdır. Dəyişikliklərin istiqamətinin proqnozlaşdırılması ilə bağlı əvvəlki ədəbiyyat əsasən səhmlərin artıq gəlirliliyi üçün zaman sıraları modellərinə əsaslanır. Məsələn, Kristofersen və Dibold (2006) müvafiq olaraq aktivlərin gəlirliliyi işarəsi proqnozu və volatillik arasındakı əlaqəni nəzərə alır və volatillik və daha yüksək şərtli momentlərin gələcək gəlirlərin işarəsini proqnozlaşdırmaqda statistik cəhətdən izahlı gücə malik olduğunu göstərir (Quantivity).

Ənənəvi olaraq bu mövzuda yuxarı və ya aşağı bir həddi aşan gəlirlilik ehtimalı əvvəlki müddətə əsaslanan məlumat əsasında şərtləndirilir. Ədəbiyyatın əksəriyyəti birdəyişənli gəlirlilik sırası ilə işləyib. Bununla birlikdə, bu ehtimalları qiymətləndirmək üçün alternativ bir yanaşma da logistik funksiyaya əsaslanaraq, multinomial logit modelindən istifadə etməkdir. Bu yanaşmanın tətbiqi ilə bağlı əsas problem izahedici dəyişənlərin seçilməsidir, çünki bu izahedici dəyişənlər ehtimalları qiymətləndirməyə kömək edən əvvəlki dövrdən təyin edilmiş məlumatların bir hissəsidir

(Quantivity).

Bu yazıda Nyu-York Fond Birjasından (NYFB) IBM səhminin tik datasından istifadə olunur. Tik data saniyələr ərzində baş vermiş hər tranzaksiyanın qeyd edilməsi ilə əldə olunan datadır. Səhmlərin gəlirliliyinin dəyişmə istiqamətini proqnozlaşdırmaq üçün cəzalandırılmış multinomial logit modelindən istifadə olunur. Xüsusilə müxtəlif dəyişiklik səviyyələrini nəzərə alaraq bir çox cavab dəyişənini (kateqoriyalar) qururuq. Modellərimizdəki cavab dəyişəni tik dataya əsaslanır və 1 baza nöqtənin (faizin yüzdə biri) həddi əsasında kateqoriyaları aşağıdakı kimi bölürük: [-2, -1], [-1,0], [0,1], [1,2]. Bu, yalnız gəlirliliyin işarəsini deyil, həm də hansı aralıqda olduğunu təxmin etməyə imkan verir.

Ehtimalların proqnozlaşdırılması üçün çox kateqoriyalı reqressiyada ən çox istifadə olunan model olduğundan multinomial logit modelinin istifadəsi seçilmişdir. Bu model dəyişənlər vektoru X-in xətti funksiyaları vasitəsilə cavab kateqoriyalarının şərti ehtimallarını müəyyənləşdirir (Tutz və Zahid 2009). Lakin dəyişənlərin sayı müşahidələrlə müqayisədə çox olduqda və ya dəyişənlər bir-biri ilə çox əlaqəli olduqda, logit modeli problemlərdən əziyyət çəkir, çünki varyasiya sonsuzluğa gedir və parametrlər qiymətləndirilə bilmir. Tənzimləmə metodlarının istifadəsi bu kimi problemləri aradan qaldırmağa kömək edə bilər. Tənzimləmə metodları cəzalandırılmış loqarifmik mümkünlük funksiyasını maksimallaşdırır. Köhnə cəza metodlarından biri olan “Ridge” (Ric) reqressiyası Şafer (1984) tərəfindən logistik reqressiya modeli üçün təyin edilmişdir. əvvəlki ədəbiyyatda çox kateqoriyalı cavablar üçün Zu və Hasti (2004) Ridge tipli cəza tətbiq etdi və Fridman (2008) L1 cəzası (LASSO), L2 (ridge) və ikisinin qarışığından (Elastic-Net) istifadə etdi.

Bu yazıda R proqramlaşdırma dilində iki fərqli paketdən istifadə edirik. MRSP paketi simmetrik tərəf məhdudiyyətləri olan çox kateqoriyalı logit model üçün Ridge reqressiyasını qiymətləndirmək üçün istifadə olunur. GLMNET paketi isə alpha

parametrini (0 dan 1-ə) dəyişməklə LASSO, Ridge və Elastic-Net qiymətləndirməsi aparır.

Tutz və Maqbul (2009) tərəfindən təqdim olunan simmetrik tərəf məhdudiyyətləri və Ridge cəzaları olan multinomial logit modeli Bölmə 2-də təqdim edilmişdir. Bölmə 3-də Fridman (2008) tərəfindən istifadə olunan və glmnet paketində mövcud olan Ridge, LASSO və Elastic-Net təqdim olunur. Bu məqalə üçün istifadə olunan data və metodologiya Bölmə 4-də təqdim olunur. Bölmə 5-də əldə edilən əsas nəticələr və fərqli modellər arasında müqayisə təqdim olunur. 6-cı bölmə yazıya yekun vurur.

Simmetrik Tərəf Məhdudiyyətli Multinomial Logit Model

Kateqoriyalı cavab dəyişəninin iki kateqoriyadan çox olması halında multinomial logit model istifadə olunur. Parametrlər qiymətləndirilə bilmədiyi üçün bəzi əlavə məhdudiyyətlərə ehtiyac var. Adətən bu məhdudiyyət tərəf məhdudiyyəti olur. Tutz və Maqbul (2009) isə alternativ məhdudiyyət təklif edirlər. Simmetrik tərəf üçün əldə edilən parametrlər ənənəvi parametrlərdən fərqlənirlər. Bu misalda median cavab referans olunan kateqoriyadır və həndəsi ortaya uyğun gəlir.Tənzimləmə metodları cəzalandırılmış loqarifmik mümkünlük funksiyasına əsaslanır.

Xüsusilə yüksək ölçülü problemlərdə tənzimləmə metodlarının istifadəsi faydalıdır, çünki cəzalandırılan qiymətləndiricilər mövcuddur və adi maksimum mümkünlük qiymətləndiricisinə nisbətən daha yaxşı proqnoz xətasına malikdir.

Glmnet paketi ilə LASSO, Ridge və Elastic-Net

Fridman (2007) ümumiləşdirilmiş xətti modelləri Elastic-Net cəzalandırma ilə uyğunlaşdırmaq üçün sürətli alqoritmlər təqdim etdi. Məqaləmizdə glmnet paketindən istifadə edərək multinomial reqressiyanı həll edirik. əvvəlki bölmədə təqdim olunandan əlavə biz burda LASSO, Ridge və Elastic-Net-i təqdim edirik.Elastic-Net cəza Ridge (a=0) və LASSO (a=1) arasında

bir yerdədir. Bu cəza xüsusilə dəyişən sayı müşahidə sayından çox olduqda və ya çoxlu bir-biri ilə yüksək korrelyasiyalı dəyişən olduqda faydalıdır.

Ridge reqressiyası əlaqəli proqnozlaşdırıcıların əmsallarına bir-birlərinə "büzüşərək"(shrinkage) bir-birlərindən güc götürməyə imkan verir. K eyni proqnozlaşdırıcıların bənzərsiz vəziyyətində hər biri vahid əmsalın 1 / k ölçüsü ilə eyni əmsallar alır (Fridman 2008).

Digər tərəfdən, LASSO çox əlaqəli proqnozlaşdırıcılara laqeyd yanaşır və birini seçib qalanlarına məhəl qoymur. Yuxarıda göstərilən unikal vəziyyətdə LASSO dağılacaqdır. LASSO cəzası bir çox əmsalın sıfıra yaxın olacağını və az sayının daha böyük və ya sıfır olmamasını gözləyir (Fridman 2008).

Əgər kiçik e> 0 üçün a = 1 – e olduqda, hər ikisinin qarışığı olan Elastic-Net LASSO-ya bənzəyir, lakin həddindən artıq korrelyasiya nəticəsində yaranan adi davranışları aradan qaldırır.

Data və Metodologiya

Məqaləmizdə NYFB-dan IBM səhminin bir aylıq tik datasını istifadə edirik. Təhlilimiz üçün məlumatları iki hissəyə bölmüşük, yəni treyninq məlumatları (ilk 3 həftə) və test məlumatları (son bir həftə). Bizə iki məlumat dəsti verildi: birində hər bir əməliyyat üçün birjanın qiyməti, digərində isə hər tranzaksiya üçün tələb və təklif qiymətləri (bid, ask prices). Hər iki məlumat dəsti müvafiq olaraq birləşdirildi və həm qiymət, həm də təklif sorğusunun mövcud olmadığı dəyərlər nəzərə alınmadı. Tələb-təklif məlumat toplusu üçün aparılan müşahidələrin ümumi sayı 287953 olmuşdur. İki məlumat toplusunu birləşdirdikdən sonra müşahidələrin ümumi sayı 74273 olmuşdur. Tələb-təklif məlumatları üçün aparılan müşahidələrin sayı əhəmiyyətlidir, çünki təhlilimizdə bizə kömək edən bir neçə izahedici dəyişəni yaratmaq üçün istifadə edilmişdir. Növbəti səhifədəki cədvəl sonrakı təhlil üçün yaradılan bütün izahedici dəyişənləri təqdim edir. Cədvəldə hər bir dəyişən

haqqında sadə bir izahat verilmişdir.

İzahedici dəyişənlər	
Dəyişən adı	İzah
Index	Hər bir müşahidə üçün gün və saatı göstərir
Date	Tranzaksiya tarixi
Time	Tranzaksiya vaxtı
Bid	Təklif qiyməti
Bidvol	Təklif həcmi
Ask	Tələb qiyməti
Askvol	Tələb həcmi
Spread	Tələb-təklif fərqi
Midprice	Tələb-təklif ortalaması
Askdif	İki tələb arasındakı fərq
Biddif	İki təklif arasındakı fərq
Askdifn	Birinci və sonrakı dördüncü tələb arasında fərq
Biddifn	Birinci və sonrakı dördüncü təklif arasında fərq
Mask	Dörd ardıcıl tələbin ortalaması
Mbid	Dörd ardıcıl təklifin ortalaması
Maskvol	Dörd ardıcıl tələb həcminin ortalaması
Mbidvol	Dörd ardıcıl təklif həcminin ortalaması
Accdiffp	Qiymət fərqinin cəmi
Accdiffvol	Həcm fərqinin cəmi
Spreadvol	Həcm fərqi
Sigmaspreadp	20 müşahidə üzrə fərqin orta kvadratik meyili
Sigmaspreadv	Həcmin orta kvadratik meyili

Sigmaask	Orta tələb qiymətinin orta kvadratik meyili
Sigmabid	Orta təklif qiymətinin orta kvadratik meyili
Dpbiddt	Ardıcıl vaxt ərzində təklif qiyməti fərqi
Dpaskdt	Ardıcıl vaxt ərzində tələb qiyməti fərqi
Dpbidvoldt	Ardıcıl vaxt ərzində təklif həcmi fərqi
Dpaskvoldt	Ardıcıl vaxt ərzində tələb həcmi fərqi

Təhlilimizdə səhm qiymətinə dair gəlirlər 1 baza nöqtəli intervallarla, yəni -1 ilə 0 bir baza nöqtəyə, 0 ilə 1 bir baza nöqtəyə və s. görə təsnif edildi. Bu intervallardan cavab kateqoriyalarının yaradılması üçün istifadə edilmişdir. Daha sağlam təhlil aparmaq üçün üç, beş, yeddi və doqquz kateqoriyadan ibarət dörd fərqli cavab kateqoriyalı dəyişənlər inşa edilmişdir.

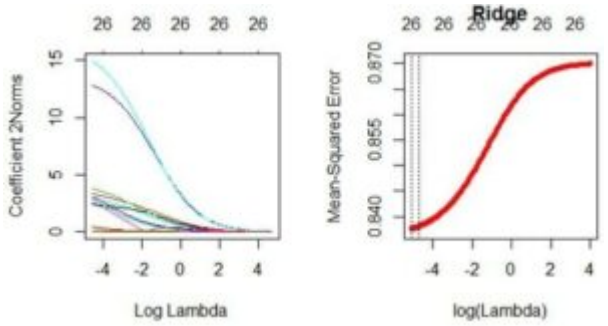
Hər bir kateqoriya səhm qiymətinin gəlirliliyinin bir baza nöqtə intervalını təmsil edir. Əvvəldə qeyd edildiyi kimi, R proqramlaşdırma dilində iki fərqli paketdən, yəni MRSP və glmnet-dən istifadə etdik. MRSP-dən istifadə edərək, simmetrik tərəf məhdudiyyət və Ridge cəzası, LASSO və ənənəvi maksimum mümkünlük təxmincisi olan multinomial logit modellər qiymətləndirildi. Buna görə, dörd fərqli cavab kateqoriyalı dəyişənlərimizlə bu paketin köməyi ilə cəmi on iki model təhlil edildi. Digər tərəfdən, glmnet alfa dəyərini 0 (Rİdge) ilə 1 (LASSO) arasında dəyişərək əlavə olaraq qarışıq cəza (Elastic- Net) əldə etməyə imkan verdi. Alfa üçün 0.1 intervalından istifadə edərək cəmi on bir müxtəlif model qiymətləndirildi. Yenə dörd cavab kateqoriyalı dəyişənlərimizlə birlikdə glmnet paketi ilə 44 fərqli qiymətləndirmə (44 model) əldə edə bildik. Bütün bu fərqli qiymətləndirmələrin köməyi ilə test məlumatlarımızda (axırıncı həftənin məlumatları) proqnozlar əldə edə bildik və dəqiqliyi qiymətləndirmək üçün bütün fərqli proqnozlarda xətaları yoxlamaq və müqayisə etməyə imkan verən bütün modellər üçün

orta kvadratik xətanı (mean-squared error, MSE) nəzərə aldıq.

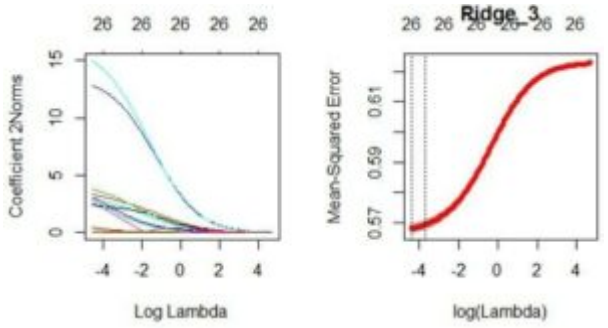
MRSP və glmnet model seçimi üçün fərqli metodlardan istifadə edir. MRSP-də model Akaike İnformasiya Kriteriyasına görə avtomatik seçilir. Digər tərəfdən, glmnet uyğun bir model seçmək üçün n (defolt = 10) qat çarpaz qiymətləndirmədən istifadə edir. Məlumatı on fərqli bloka bölməklə bunu edir və modeli qurmaq üçün doqquz blokdan istifadə edir və qalan blok üçün orta kvadratik səhvləri əldə edir. Nəticə etibarı ilə orta kvadratik xətanın minimuma endirildiyi lambdanı seçir. Glmnet-dəki cv.glmnet funksiyası bizə MSE (orta kvadratik xəta)- lambda üçün də qrafik çıxarmağa imkan verir. Hər iki paketin bir çatışmazlığı odur ki, xüsusilə glmnet vəziyyətində alfalar üçün döngələri işə salmaq üçün xeyli vaxt gedir. Bununla birlikdə, hər iki paketin istifadə edilməsinin üstünlüyü müəyyən bir istinad kateqoriyasına ehtiyac duymadıqları, əksinə modelləri qiymətləndirmək üçün simmetrik multinomial logit modelindən istifadə etmələridir.

Sonrakı nəticələrdə göstərildiyi kimi, burada LASSO əmsalların əksəriyyətini kənara qoymağa meyillidir, çünki bütün yüksək əlaqəli əmsalları sıfıra endirir. Təhlilimizdəki əmsalların əksəriyyəti yüksək əlaqəyə malikdir, çünki hamısı eyni tələb-təklif data dəstindən qurulmuşdur.

Bəzi hallarda LASSO çox əhəmiyyət kəsb etməyən sabit dəyişəndən başqa bütün əmsalları sıfıra endirir, digər tərəfdən Ridge və Elastic-Net dəyişənlərin əksəriyyətini azaltmağa meyillidir, lakin sıfıra endirmir. Buna baxmayaraq, bütün təhlil zamanı məqsədimiz izahedici dəyişənləri seçməkdən daha yüksək proqnozlaşdırıcı gücə malik modeli seçmək idi. Bunun səbəbi, əsas məqsədimiz, dataya gələcək gəlirliliyi mümkün qədər səmərəli şəkildə proqnozlaşdıra biləcək bir məlumat toplusu kimi yaxınlaşmaqdır. Bu səbəbdən ən aşağı orta kvadratik xətalı (MSE) model ilə maraqlanırıq. Intuitiv olaraq cavab kateqoriyasının sayı azaldıqca, MSE-nin azalacağını gözləməliyik, amma təhlillərimiz göstərir ki, bəzən belə olmur.



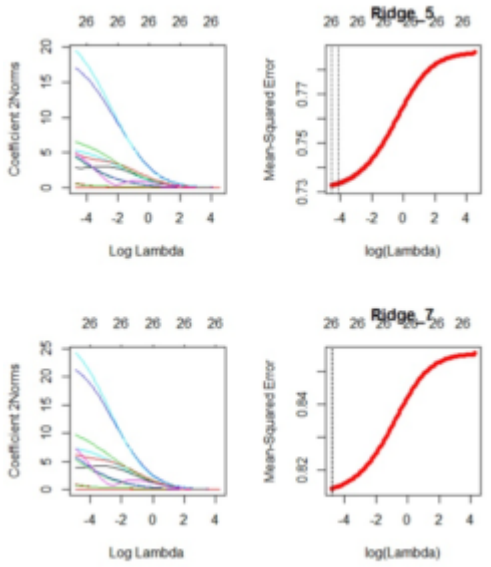
Yuxarıdakı qrafiklər cv.glmnet-dən doqquz istinad kateqoriyası olan modeldən istifadə edərək əldə edilən nəticədir. Sol tərəfdəki qrafik lambda, yəni tənzimləmə parametrinin fərqli dəyərləri üçün əmsal yollarını göstərir. Bu vəziyyətdə əmsallar L-2 normasındadır (Evklid norması). Sağdakı qrafik 10 qat çarpaz qiymətləndirmədən istifadə edilərək Ridge cəzası olan modeli göstərir. Sağdakı qrafikdə iki nöqtəli xətt var. Solda olan MSE-nin minimum olduğu lambda, sağdakı isə minimum MSE-nin standart xətasına uyğun lambdanı verir. Təhlildəki məqsədimiz təsdiq edilmiş MSE-ni minimuma endirən lambda əsasında əmsallar və model seçməkdir.



Yuxarıdakı qrafiklər üç kateqoriyalı keysə aiddir. Bu keysdəki əmsal yolları doqquz kateqoriyalı keys ilə oxşar görünür, lakin sağ tərəfdəki qrafik daha uyğun model olduğunu göstərir (daha az xətalıdır). İki nöqtəli xətt bir-birinə daha yaxındır və bu daha yaxşı uyğunluğun əksidir. Bu, müəyyən mənada intuisiyamızı sübut edir və daha az kateqoriyalar daha yaxşı uyğunlaşmaya (better fit) səbəb olur.

Yuxarıdakı qrafiklər digər iki, yəni beş və yeddi kateqoriyalı keyslərə aiddir. Daha az kateqoriyalarda daha yaxşı uyğunlaşma meyili də bu qrafiklərdən müəyyən dərəcədə aydındır. Bununla birlikdə, minimum MSE-lərin paylanmasında erkən intuisiyaya

qarşı gedən fərq var.



Yuxarıdakı dörd qrafik üç cavab kateqoriyası olan model üçündür. LASSO, alfa = 0.5 olan Elastic-Net cəzadan daha yaxşı çıxış edir. LASSO 27 dəyişəndən 11-ni sıfıra, Elastic-Net isə 9 dəyişəni sıfıra endirir. Ridge cəzası hər hansı bir dəyişəni sıfıra endirməsə də, LASSO və Elastic- Net belə deyil. Bu, təhlilimiz üçün alfanın düzgün seçiminin vacibliyini göstərir. Bu səbəbdən biz cv.glmnet funksiyasında 0-dan 1-ə (0.1 intervalla) qədər olan 11 fərqli alfa dəyərindən istifadə edirik və daha yaxşı uyğunlaşma və proqnozlaşdırıcı güc (daha kiçik xəta) ilə modeli seçirik.

Nəticələr və Analiz

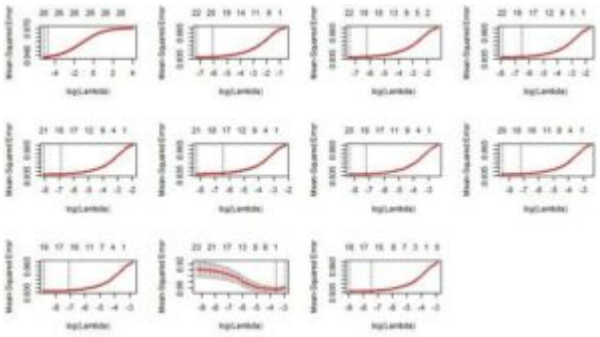
Daha əvvəl qeyd edildiyi kimi cv.glmnet funksiyası ilə təchiz olunmuş modellərin çıxışını təhlil edirik. Dörd fərqli cavab kateqoriyası var idi, yəni 3, 5, 7 və 9 cavab kateqoriyası.

9 Kateqoriyalı Data	
L1 -in Çəkisi	Proqnozlaşdırılan MSE
α = 0.0	2.67258
α = 0.1	2.72469
α = 0.2	2.72583
α = 0.3	2.73232

α = 0.4	2.73291
α = 0.5	2.73329
α = 0.6	2.73248
α = 0.7	2.73383
α = 0.8	2.73194
α = 0.9	3.68399
α = 1.0	2.73746

Yuxarıdakı cədvəldə 0-dan 1-ə qədər olan fərqli alfa dəyərlərində 9 kateqoriyalı data dəsti üçün proqnozlaşdırılan MSE-lər verilmişdir, burada alfa = 0 Ridge cəzası və alfa = 1 LASSO -dur. Cədvəldən ən aşağı MSE-nin Ridge təxmincisindən nəticələndiyini görə bilərik (alfa = 0). Bundan belə nəticə çıxara bilərik ki, bəzilərini sıfıra endirmək əvəzinə, modeldəki dəyişənlərin hamısını (27-sini də) saxlamaq daha yaxşıdır. MSE-lər sinif (kateqoriya) proqnozlarından istifadə etməklə əldə edilir.

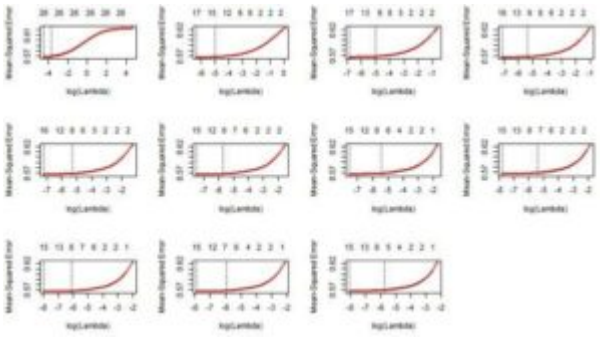
Aşağıdakı qrafiklər dəsti, 0-dan 1-ə qədər dəyişən alfa qiymətləri üçün 9 kateqoriyalı modellərdən çarpaz qiymətləndirmə yollarını təqdim edir. Yollar olduqca oxşar görünsə də, Ridge qiymətləndiricisi olan model (ilk qrafik) minimum MSE üçün nisbətən daha yaxşı paylanmanı təmin edir. Bu o deməkdir ki, digər modellərlə müqayisədə MSE-ni minimuma endirən lambda dəyəri və minimum MSE-nin bir standart xətası daxilində lambdanın dəyəri çox yaxındır. Bu, həqiqətən daha uyğun bir model üçün yaxşı bir əlamətdir. Qrafiklərdəki MSE-lər çarpaz doğrulama proqnozlarından əldə olunduğundan, yalnız kvadratik proqnoz xətalarını göstərir. Buna görə MSE-lərin paylanmasını (bu halda minimum MSE) yoxlamaq daha ağlabatan bir fikirdir.



Cavab dəyişənlərini yoxlamaq və müqayisə etmək üçün indi digər üç kateqoriyalı modelləri təqdim edirik.

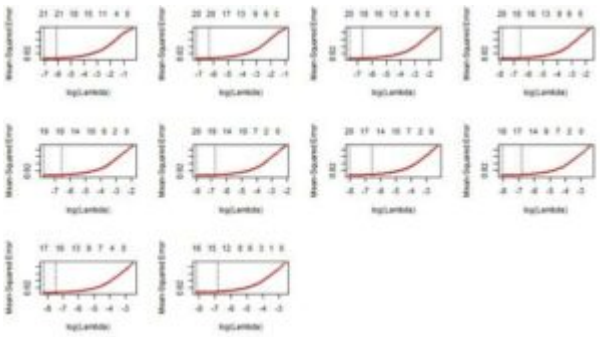
3 Kateqoriyalı Data	
L1 -in Çəkisi	Proqnozlaşdırılan MSE
α = 0.0	2.77501
α = 0.1	2.78172
α = 0.2	2.78237
α = 0.3	2.78410
α = 0.4	2.78367
α = 0.5	2.78259
α = 0.6	2.78280
α = 0.7	2.78367
α = 0.8	2.78454
α = 0.9	2.78064
α = 1.0	2.78324

Yuxarıdakı cədvəldə 3 cavab kateqoriyası olan model üçün proqnozlaşdırılan MSE-lər verilmişdir. Yenə buna bənzər bir yanaşma istifadə edərək, Ridge qiymətləndiricisinin digər cəzalara nisbətən daha yaxşı hərəkət etdiyini görürük. Burada 9 cavab kateqoriyası olan modellə müqayisədə bu modelin daha aşağı model MSE-lər, lakin daha pis proqnozlaşdırılan MSE-lər verdiyini qeyd etmək vacibdir. Aşağıdakı qrafiklər dəsti fərqli alfa dəyərləri üçün 3 cavab kateqoriyalı modellərdən çarpaz qiymətləndirmə yollarını təqdim edir.



Eynilə, cədvəl və aşağıdakı qrafiklər dəsti 7 cavab kateqoriyası olan modellər üçün müvafiq olaraq proqnozlaşdırılan MSE-lər və çarpaz qiymətləndirmə yollarını təqdim edir.

7 kateqoriyalı data	
L1 -in çəkisi	Proqnozlaşdırılan MSE
α = 0.0	2.68606
α = 0.1	2.72107
α = 0.2	2.72328
α = 0.3	2.72945
α = 0.4	2.73140
α = 0.5	2.72583
α = 0.6	2.72718
α = 0.7	2.72777
α = 0.8	2.72761
α = 0.9	2.72945
α = 1.0	2.73237

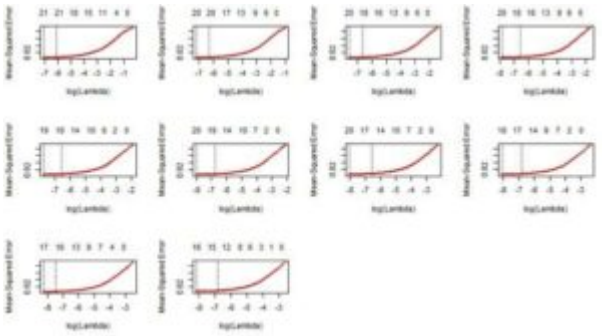


Əvvəlki iki vəziyyətə bənzər şəkildə, Ridge qiymətləndiricisi yeddi cavab kateqoriyası ilə olan modeldə qalanlarını

üstələyir. Hərçənd, bu vəziyyətdə minimum MSE paylanmalarının alfa-nın fərqli dəyərləri arasında fərqli olub olmadığı məlum deyil. Proqnozlaşdırılan orta kvadratik xətaya gəlincə, 7 cavab kateqoriyası ilə olan model 3 cavab kateqoriyası ilə olan modeli üstələyir, lakin 9 cavab kateqoriyası ilə olan modellə müqayisədə daha pisdir.

Aşağıdakı cədvəl və qrafik dəsti 5 cavab kateqoriyası ilə olan son modelimiz üçün müvafiq olaraq proqnozlaşdırılan MSE və çarpaz qiymətləndirmə yollarını təmin edir.

5 Kateqoriyalı Data	
L1-in Çəkisi	Proqnozlaşdırılan MSE
α = 0.0	3.00368
α = 0.1	3.00660
α = 0.2	3.00628
α = 0.3	3.00179
α = 0.4	3.00281
α = 0.5	3.00400
α = 0.6	3.00579
α = 0.7	3.00525
α = 0.8	3.00617
α = 0.9	3.00676
α = 1.0	3.00801



Yuxarıdakı nəticələr göstərir ki, alfa = 0.3 (Elastic-Net) olan modelin qalan hissədən proqnozlaşdırıcı gücü üstündür. Yenidən, çarpaz doğrulama yollarında fərqli süjetləri ayırd

etmək üçün sübut yoxdur. Alfa = 0.3 üçün MSE-nin ən aşağı qiymətini nəzərə alsaq, bu vəziyyətdə bu modelin 5 cavab kateqoriyası ilə olan modellərin ən yaxşısı olduğu təxmin edilə bilər. Digər 3 model dəsti ilə müqayisədə 5 cavab kateqoriyası ilə olan modellər ən yüksək proqnoz səhvlərini verir.

MRSP

Glmnet paketindən əldə edilən nəticəni təhlil etdikdən sonra diqqətimizi bu iş zamanı istifadə edilən digər paketə, yəni MRSP-yə yönəldirik. MRSP simmetrik tərəf məhdudiyyətləri və Ridge cəzası ilə multinomial logit modelini qiymətləndirmə seçimini verir. Model, hər bir kateqoriyaya münasibətdə hər bir test müşahidəsi üçün proqnozlaşdırılan ehtimalları təmin edir. Bu ehtimalları əldə etdikdən sonra gözlənilən sinif bu ehtimalları istinad sinifinə vurmaqla əldə edilir. Bu, proqnoz səhvlərini, yəni proqnozlaşdırılan orta kvadratik xətanı hesablamağa imkan verir.

Bundan əlavə, eyni model üçün LASSO və sadə multinomial logit ilə də proqnoz xətalarını hesablayırıq (cəzasız və istinad kateqoriyası = 0 olmaqla). Bu, Ridge qiymətləndiricisi ilə müqayisə etmək və səmərəliliyin əldə olunub-olunmadığını anlamaq üçün edilir. Bundan əlavə, tənzimləmə parametri, lambda, AIC (Akaike İnformasiya Kriteriyası) və BIC (Bayes İnformasiya Kriteriyası) istifadə edərək seçilir. Aşağıdakı iki cədvəl müvafiq olaraq AIC və BIC istifadə edərək proqnozlaşdırılan MSE-ləri təqdim edir.

AIC			
Kateqoriya	Proqnozlaşdırılan MSE – LASSO	Proqnozlaşdırılan MSE – Cəzasız	Proqnozlaşdırılan MSE – Simmetrik Ridge
3	2.52297	2.52256	2.52182
5	2.47749	2.47409	2.47339
7	2.52254	2.51958	2.51893

9	2.80025	2.80066	2.80217

Kateqoriya	Proqnozlaşdırılan MSE – LASSO	Proqnozlaşdırılan MSE – Cəzasız	Proqnozlaşdırılan MSE – Simmetrik Ridge
3	2.52499	2.52256	2.52182
5	2.48389	2.47409	2.47339
7	2.52513	2.51958	2.51893
9	2.79353	2.80066	2.80217

Yuxarıda göstərilən iki cədvəldən istifadə edib deyə bilərik ki, simmetrik Ridge digər iki modeli istifadə olunan informasiya meyarından asılı olmayaraq üstələyir. Bununla birlikdə 9 cavab kateqoriyası ilə olan model üçün simmetrik Ridge qiymətləndiricisi həqiqətən ən pisini yerinə yetirir. Bu, bu vaxta qədər aşkarladığımız çox şeyə qarşı çıxır.

Nəticə

Yuxarıda göstərilən təhlillərdən belə nəticəyə gəlmək olar ki, bir neçə hal istisna olmaqla, Ridge qiymətləndiricisi ən aşağı orta kvadratik xətaya nail olmaqda digər qiymətləndiricilərdən üstündür. Daha az sayda kateqoriyaların daha yaxşı bir modelə sahib olması ilə səciyyələnən əvvəlki intuisiyamızın proqnozlaşdırılan MSE-lər tərəfindən deyil, model MSE-lər tərəfindən dəstəkləndiyini öyrəndik. Təhlilimiz üçün yalnız bir aylıq data var idi. 3 həftəlik datanı treyninq etdik və son bir həftəni testdən keçirdik. Datamızın təbiətinə görə ola bilər ki, biz nəticəyə zidd olan bəzi anomaliyalar əldə edirik. Ancaq nəticələrin əksəriyyəti Ridge qiymətləndiricisinin ən yaxşı proqnozlaşdırma gücünə sahib olduğunu göstərir və bu nəticə Tutz və Maqbul (2009) tərəfindən hazırlanan məqaləyə uyğun gəlir. Bənzər bir analiz daha böyük bir data dəsti üzərində və ya daha kiçik (həftəlik, gündəlik, güniçi) intervallarla aparılarsa, proqnozlaşdırılan gücün yaxşılaşdığı hal ola bilər. Əlavə təhlil üçün kateqoriyaların sayını artırmaq bir seçim ola bilər, ancaq paketlərin məhdud hesablama gücü ilə analiz baş tutmaya bilər.

Buna baxmayaraq, istifadə olunan müxtəlif texnikalara əsaslanaraq, simmetrik tərəf məhdudiyyətləri və Ridge cəzası olan multinomial logit model vaxt səmərəliliyi və proqnoz dəqiqliyi baxımından ən yaxşısını həyata keçirdi.

Ədəbiyyat siyahısı:

Christoffersen et al, 2004. Direction-of-Change Forecasts Based on Conditional Variance, Skewness and Kurtosis Dynamics: International Evidence

Friedman, J., Hastie, T., Tibshirani, R., 2008. Regularization paths for generalized linear models via coordinate descent.

Nyquist, H., 1991. Restricted estimation of generalized linear models. Journal of Applied Statistics 40, 133–141.

Schaefer, R., Roi, L., Wolfe, R., 1984. A ridge logistic estimator. Communications in Statistics: Theory and Methods 13, 99–113.

Tibshirani, R., 1996. Regression shrinkage and selection via lasso. Journal of the Royal Statistical Society B 58, 267–288. https://quantivity.wordpress.com/2012/01/16/sign-direction-of-change-forecasting/

Tutz, Gerhard, Zahid, M. Faisal, 2009. Ridge Estimation for Multinomial Logit Models with Symmetric Side Constraints. http://www.stat.uni-muenchen.de/

Zhu, J., Hastie, T., 2004. Classification of gene microarrays by penalized logistic regression. Biostatistics 5, 427–443.

Qeyd: Bu məqalə Konstanz Universitetində Daniyal Rizvan ilə birlikdə Big Data seminarı üçün ingilis dilində yazılmışdır.