

Kredit riskinin modelləşdirilməsi: PD modelləşdirməsində daralma üsulları və LASSO seçimi (1- ci hissə)

written by Hikmat Abdulazizov Hikmət Əbdüləzizov

Maliyyə sənayesinin dinamik mühitində kredit riskinin dəqiq qiymətləndirilməsi kredit təşkilatlarının sabitliyinin və dayanıqlığının qorunmasında əsas amil kimi çıxış edir. Qlobal iqtisadiyyat davamlı olaraq inkişaf etdikcə, kredit risklərinin idarə edilməsi ilə bağlı mürəkkəblilər və problemlər də inkişaf edir. Bu kontekstdə mütərəqqi ekonometrik üsullardan və müxtəlif verilənlər bazalarından istifadə kredit riski modellərinin dəqiqliyini və etibarlılığını qorumağa çalışan maliyyə institutları üçün vacibdir. Kredit riski, borc alanın öz maliyyə öhdəliklərini yerinə yetirə bilməməsi potensialına aiddir və bu, borc verən üçün maliyyə itkisinə səbəb olur. Bu, mahiyyətcə borc alanların kreditlərini ödəyib-ödəməmələri və/və ya müqavilə öhdəliklərini yerinə yetirib-yetirməmələri ilə bağlı qeyri-müəyyənlikdən yaranan riskdir. Kredit riskinin effektiv idarə edilməsi maliyyə institutlarının sabitliyini və ödəmə qabiliyyətini qorumaq və sağlam maliyyə sistemini saxlamaq üçün çox vacibdir. Kredit riskinin modelləşdirilməsi sahəsi kredit riskini proqnozlaşdırmaq və kəmiyyətləndirmək üçün mürəkkəb statistik və riyazi modellərin işlənilməsi, kreditövlərə kredit verməklə bağlı əsaslandırılmış qərarlar qəbul etməyə və ümumi riskə məruz qalmalarını idarə etməyə kömək etməyi əhatə edir. Qabaqcıl statistik və maşın öyrənmə üsullarının bu cür modellərə inteqrasiyası kredit qabiliyyətinin daha ətraflı qiymətləndirilməsinə imkan verir ki, bu da maliyyə

institutlarına əsaslandırılmış kredit qərarları qəbul etməyə və kapitalın bölüşdürülməsini optimallaşdırmağa imkan verir.

Defolt

Bank işində kredit riski kontekstində 90 gün gecikmə meyarı ilə defolt, borc alan öz kredit və ya kredit öhdəliyi üzrə ödənişini 90 gündən çox müddətə yerinə yetirmədikdə (Azərbaycan bazarı və ya ona oxşar bazarlarda ardıcıl olaraq 3 ay ödəmədikdə) baş verir. Bu, gecikmədə olan kredit müqaviləsinin əhəmiyyətli dərəcədə pozulması hesab edilir və kredit riskinin yüksək səviyyəsini göstərir. Maliyyə institutları adətən 90 gün və daha artıq gecikməsi olan borc alanları defolt kimi təsnif etmək üçün əhəmiyyətli mərhələ kimi istifadə edirlər.

Defolt Ehtimalının (PD) dəqiq ekonometrik qiymətləndirilməsi kredit riskinin qiymətləndirilməsində mühüm komponentdir. Son illərdə PD modellərinin dəqiqliyini və sabitliyini əldə etmək üçün büzülmə üsullarının və Lasso seçim üsullarının gücü praktikada daha çox qəbul olunur. Ənənəvi PD modelləşdirməsi adətən logistik reqressiya kimi ənənəvi statistik üsulları əhatə edir. Logistik reqressiyada proqnozlaşdırıcı dəyişənlər (borc alanın kredit qabiliyyəti ilə bağlı xüsusiyyətlər) və ikili hədəf nəticə (default və ya qeyri-defolt) arasında əlaqə modelləşdirilir. Logistik reqressiyadan əldə edilən əmsallar hər bir proqnozlaşdırıcının defolt ehtimalına təsirini əks etdirir. Bu modelin üstünlükləri onun əmsallarının asan şərh edilməsi və sadəliyidir. Logistik reqressiya yaxşı qurulmuş və geniş istifadə olunan bir üsuldur. Bununla belə, bu, ən qədim təsnifat modeli ola bilər.

Başqa bir problem ondan ibarətdir ki, logistik reqressiya proqnozlaşdırıcılar və defoltun log-əmsalları arasında xətti əlaqəni nəzərdə tutur və o, verilənlərdə mürəkkəb, qeyri-xətti əlaqələri tutmaya bilər. Azərbaycanın hazırki bank sektorunda hətta ən qabaqcıl banklarında belə hələ də yalnız bu modeldən istifadə edilən keyslər vardır. Üstəlik, mühüm dəyişənlərin

seçimi bu modelle olduqca ağır olur. Təsəvvür edin ki, sizdə 10 dəyişən var. Sonra $2^{(10)}=1.024$ mümkün modelə sahib olmalısınız. Əgər 20 xüsusiyyətiniz varsa, o, 1.048.576 mümkün modelə çevrilir ki, bunu da məhdud zaman intervalında hesablamaq mümkün deyil. Məqalənin 2-ci hissəsində bu problemlə məşğul olacağıq.

Bundan əlavə, büzülmə texnikasının bir variantı olan Lasso seçimi xüsusiyyət seçiminə və modelin şərh edilməsinə kömək edir. Model əmsallarında seyrəkliyə səbəb olmaqla, Lasso avtomatik olaraq ən uyğun dəyişənləri müəyyən edir və bu, daha azdəyişənli və şərh edilə bilən modellərə gətirib çıxarır. Bu metod atribut kimi kredit riskinin modelləşdirilməsində böyük əhəmiyyət kəsb edir, çünki o, maliyyə institutlarına defolt ehtimalını şərtləndirən əsas amilləri dəqiq müəyyən etməyə imkan verir.

Bu çox vacib bir mövzudur, çünki bu hesablamalar son nəticədə bankın mənfəətinə birbaşa təsir edir. Bundan əlavə, LASSO yanaşması təkcə banklarda deyil, həm də ağlabatan xüsusiyyət seçimi və proqnozlaşdırılmasına ehtiyacı olan müxtəlif sənaye sahələrində tətbiq oluna bilər. İstifadə etdiyimiz məlumatlar [açıq mənbədən](#) götürülüb və 307.512 fiziki şəxs (və ya kredit) və 122 xüsusiyyətdən ibarətdir. Beləliklə, verilənlər bazası çox zəngindir və bu tədqiqatda istifadə etməklə dəstəklənən metodlar nəinki Azərbaycanın istənilən bankına tətbiq oluna bilər və ya tətbiq edilməlidir, həm də məlumat toplusu Azərbaycanın maliyyə bazarında nələrin məlumat kimi qeyd edilməli olduğu barədə fikir verir. Bu bazar, məlumatlar və modelləşdirmə mədəniyyəti baxımından hələ də primitiv inkubator vəziyyətindədir.

İki məqalədən ibarət olacaq məqalələr silsiləsi bazar və tədqiqatçılar üçün əhəmiyyətli dərəcədə faydalı metodologiya və yanaşma təqdim edəcəkdir. Birinci məqalədə biz bir qədər adi modelləşdirmə yanaşmasını və onun çətinliklərini *Ridge of büzülmə* modelləşdirməsi ilə birlikdə göstərəcəyik. İkinci məqalədə biz Robert Tibşiraninin (Tibshirani 1996) LASSOizmini

dərindən araşdıracağıq. Beləliklə, müasir texnikalardan istifadə etməklə vaxtı və mürəkkəbliyi nə qədər səmərəli şəkildə azalda biləcəyimizi başa düşəcəyik.

Metodologiya

Yuxarıda qeyd etdiyimiz kimi, burada istifadə olunan verilənlər bazası açıq mənbədir və təxminən 300.000 fərddən və onların zəngin xüsusiyyətlərindən ibarətdir. Məlumatların təqribən 10%-i 1-dir (1=defolta düşmüş şəxs). Beləliklə, biz hədəflənmiş defolt nəticəni proqnozlaşdırmaq üçün nədən istifadə etməliyik? (Hədəflənmiş defolt dəyişən 1-lər və 0-lərdir.) Cavab olduqca sadədir: əlimizdə nə varsa və “daha çox”. Buradakı “daha çox” o deməkdir ki, bizdə mövcud olanlardan daha çox funksiya yarada bilərik. Məsələn, yaş məlumatımız varsa, qeyri-xəttiliyi nəzərə alan daha çox dəyişən yaratmaq üçün onu kvadrat və ya kub edə bilərik. Bu cür dəyişənlərə monoton əlaqələr yaratmaq üçün sübut ağırlığından (WOE) istifadə etməyi tövsiyə etmirik. Aşağıda datanı daha aydın başa düşməyiniz üçün istifadə edilən məlumatların çox kiçik bir hissəsi verilmişdir. Ümumiyyətlə, banklarda ekonometrik təhlil üçün oxşar məlumatlar var. Ən vacib və qiymətli məlumatlar, əlbəttə ki, gecikmə pozuntuları ilə bağlı məlumatlardır.

Cədvəl 1. Xüsusiyyətlər

ID	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	AGE	FLAG_OWN_CAR
100002	1	Cash loans	M	43	N
100003	0	Cash loans	F	26	N
100004	0	Revolving loans	M	34	Y
100006	0	Cash loans	F	63	N

Burada biz yalnız nağd kreditlərdən istifadə edirik, çünki dövri kreditlərə (kredit kartlarına) ayrıca müraciət etmək lazımdır. Biz məlumatları treyninq dəstinə, validasiya dəstinə və test dəstinə ayırırıq. Hiperparametri, yəni kəsməni tənzimləmək üçün bizə validasiya dəsti lazımdır. Bizim

logistik regressiyamız defolt ehtimallarını proqnozlaşdırdığı üçün kəsime ehtiyacımız var və onu test dəstində sınaqdan keçirmək üçün kəsmə parametrindən (sıfırla bir arasında) istifadə edərək onları sıfırlara və birlərə çevirməliyik. Biz adi logistik regressiyanın vaxt baxımından effektiv olması üçün dəyişənlərin geriyə doğru seçilməsindən istifadə edirik. 100-dən çox xüsusiyyətdən 20 dəyişənlə yekunlaşırıq. Daha az dəyişənə sahib olmaq üçün xüsusiyyətlərin seçilməsi üçün 1%-dən az p-dəyərlərindən istifadə etdik. Əlbəttə ki, 5%-dən az olan p-dəyərlərindən istifadə etsək, daha çox nəticə əldə edə bilərik.

P-dəyəri statistik testlə əlaqəli bir ehtimaldır. 0, boş bir fərziyyəyə qarşı sübutları kəmiyyətləşdirir. Fərziyyə testində sıfır fərziyyə heç bir təsirin və ya heç bir fərqi olmadığını dair defolt fərziyyəni təmsil edir. P-qiyməti sıfır fərziyyə doğru olduqda müşahidə edilən nəticələrin və ya daha ekstremal nəticələrin əldə edilməsi ehtimalını göstərir. Əgər p əvvəlcədən müəyyən edilmiş əhəmiyyət səviyyəsi alfa-dan azdırsa və ya ona bərabərdirsə: Sıfır fərziyyəni rədd edin və dəyişənin əhəmiyyətli olduğunu tapın. Əgər $p > \text{alfa}$: sıfır fərziyyəni rədd edə bilmirsinizsə, o, əhəmiyyətsiz dəyişəndir. Yaş dəyişkənliyini bir az daha vurğulamaq istəyirik. $0 (a \cdot x - b \cdot x^2)$ şəkliyədir. Bu, aşağıya doğru paraboladır. Bu o deməkdir ki, bizim standart cavab dəyişənimiz yaşla qeyri-xətti əlaqəyə malikdir. Müəyyən bir nöqtədə, defolt ehtimal yaşla birlikdə artır və bir müddət sonra azalır.

Bu müşahidə qazancların orta yaşda maksimuma çatması ilə bağlı məlum faktla bağlı ola bilər. Buna görə də bu növ dəyişənlər üçün sübutların çəkisi kimi xətti çevrilmələri tövsiyə etmirik.

Sonra validasiya dəstindən istifadə edərək hiperparametrin kəsilməsini təsdiq etməliyik. Treyninq məlumat dəsti ilə 0 və 1 arasında defolt ehtimallarını proqnozlaşdırdıqdan sonra defolt və ya qeyri-defolt kimi nə adlandıracağımızı bilmədiyimiz üçün bu əhəmiyyətli məsələdir. Sinifləri 0 və ya

1 kimi proqnozlaşdıran funksiyadan istifadə oluna bilər. Lakin bu, proqrama ehtimallar üçün kəsmə dəyərini 0,5 kimi seçməyə məcbur edəcək. Bu da öz növbəsində həddindən artıq sadələşdirməyə gətirib çıxardığına görə yaxşı əsaslandırılmamış olur. 0,5-lik kəsim dəyəri o deməkdir ki, 0,5-dən yuxarı ehtimallar 1, əks halda 0 kimi proqnozlaşdırılır.

Apriori normal görünərsə də, belə bir kəsimin əsaslandırılması üçün hər bir məlumat dəstini yoxlamaq lazımdır. Burada kəsmə dəyərlərini əldə etdiyimiz meyarları təqdim etməli və sonra modelləri müqayisə etməliyik. Biz AUC (əyri altındakı sahə) istifadə edə bilərik, lakin modelləri təsadüfi meşə və ya qərar ağacı kimi maşın öyrənmə analoqları ilə müqayisə etmək üçün daha universal meyarlara ehtiyacımız var.

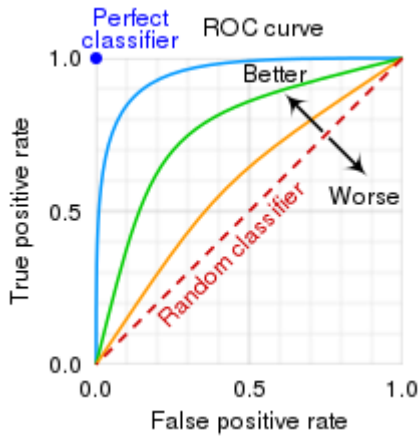
Üç anlayışı müəyyən etmək lazımdır: dəqiqlik, həssaslıq və spesifiklik. Dəqiqlik modelin düzgün tutduğu hissədir. Dəqiqlik 90%-dirsə, sıfırların və birlərin 90%-i düzgün proqnozlaşdırılıb. Həssaslıq modelin həqiqi müsbət (TP) nisbətlərini (1s) düzgün proqnozlaşdırma sürətini ölçür. Əksinə, spesifiklik isə modelin neqativləri (0s) düzgün proqnozlaşdırma sürətini ölçür.

Qrafik 1. Qarışıqlıq Matrisi

		Predicted Value	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Üç modelin hər biri üçün kəsmə dəyərlərini seçmək üçün üç meyar istifadə olunur. Birincisi, dəqiqliyi maksimuma çatdıran kəsmədir. İkincisi, ROC əyri qrafikinə yuxarı sol küncü ilə əyrinin özü arasındakı məsafəni minimuma endirərək həm həssaslığı, həm də spesifikliyi artırır. 0,5-ə bərabər olan əyri altındakı sahə heç bir proqnozlaşdırıcı gücü olmayan təsadüfi təsnifatı göstərir. AUC əvəzinə Gini istifadə etmək də tövsiyə edilmir.

Qrafik 2. ROC əyrisi

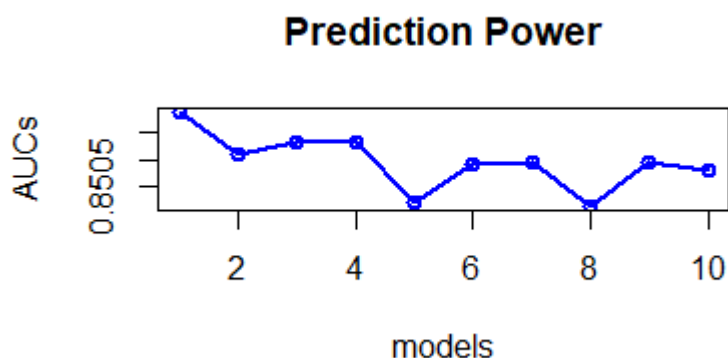


Sonuncu meyar, öz-özünə müəyyən edilmiş xərc funksiyasının minimuma endirilməsidir. Bu xərc funksiyası yalan neqativləri və yalan müsbətləri ümumiləşdirir ki, yanlış neqativlər yalançı pozitivlərdən üç dəfə baha başa gəlir. Bu halda, bu o deməkdir ki, defoltun qeyri-defolt kimi yanlış etikətlənməsinin dəyəri qeyri-defoltun defolt kimi yanlış etikətlənməsindən qat-qat baha başa gəlir, çünki birincisi daha risklidir. Kəsmə hiperparametrinin böyük əhəmiyyəti var, çünki o, maliyyə institutunun, bankın mənfəətinə birbaşa təsir göstərir.

Burada büzülmə modelləri dünyasına daxil oluruq. 2-ci hissədə biz onları daha yaxşı tədqiq edəcəyik, lakin burada adi logistik reqressiya ilə modelləri müqayisə etmək üçün onlara ehtiyacımız var. Reqressiya və ya cəzalandırılmış reqressiya modelləri kimi də tanınan büzülmə modelləri, reqressiya əmsallarına cəza həddi əlavə etməklə həddən artıq uyğunlaşma və multikollinearlıq kimi məsələləri həll etmək üçün istifadə edilən statistik üsullardır. Bu modellər təxmin edilən əmsalları sıfıra doğru azaltmaqla və ya nizamlamaqla modelin sabitliyini və ümumiləşdirmə performansını yaxşılaşdırmaq məqsədi daşıyır. İki ümumi büzülmə modelləri Ridge və Lassodur. Onların modeldə əlavə qərəzlilik tətbiq etməklə ümumi orta kvadrat səhvini yaxşılaşdırdıqları da məlumdur. Cəzalandırılmış logistik reqressiya hallarında büzülmə modellərinin tənzimlənməsində alfa parametri (sıfırla bir arasında) mövcuddur. 0 ilə 1 arasında sonsuz sayda həqiqi ədədlər olduğundan, biz hər dəfə alfa parametrini 0,1

artırmaqla onlardan yalnız 11-ni istifadə edirik və modelləri müvafiq AUC-lərlə (ROC əyrisi altındakı sahə) müqayisə edirik. Qısa desək, biz burada təmiz maşın öyrənmə modellərindən istifadə etmirik.

Qrafik 3. Modellər və Proqnoz Gücləri



Gəlin onları bir modelə endirmək üçün seçilmiş modellərin nəticələrinə baxaq.

Cədvəl 2. Modellərin Nəticələri

MODEL 2 alpha = 0.1					
Accuracy Cutoff	0.32	Cost Minimum Cutoff	0.24	Minimum Distance Cutoff	0.09
Accuracy	0.83	Accuracy	0.81	Accuracy	0.76
Sensitivity	0.21	Sensitivity	0.29	Sensitivity	0.73
Specificity	0.93	Specificity	0.88	Specificity	0.67
MODEL 6 alpha = 0.6					
Accuracy Cutoff	0.35	Cost Minimum Cutoff	0.19	Minimum Distance Cutoff	0.11
Accuracy	0.88	Accuracy	0.83	Accuracy	0.73
Sensitivity	0.25	Sensitivity	0.27	Sensitivity	0.76
Specificity	0.96	Specificity	0.83	Specificity	0.62
MODEL 8 alpha = 0.8					
Accuracy Cutoff	0.43	Cost Minimum Cutoff	0.22	Minimum Distance Cutoff	0.07
Accuracy	0.87	Accuracy	0.79	Accuracy	0.79
Sensitivity	0.32	Sensitivity	0.31	Sensitivity	0.71
Specificity	0.96	Specificity	0.81	Specificity	0.72

Yuxarıdakıların hamısı Lasso və Ridge arasında yerləşən elastik şəbəkə büzülmə modellərini təsvir edir (alfa 0 və 1 dəyərlərini dəqiq qəbul etmir, lakin onların arasındadır). Mən bunları bir bankın menecerlərinə təqdim edərdim və kəsimlər arasında variantlar təqdim edərdim çünki bu, onlara da seçim imkanı verir. Bununla belə, texniki dildə desək, ən yüksək həssaslığa malik minimum məsafə kəsimini seçərdim, çünki bu, ən güclüdür. Modelimizi seçdikdən sonra, əlbəttə ki, test verilənlər bazasında sınaqdan keçirə bilərik. Sonra, gələcək defoltrları tapmaq və gözlənilən kredit itkimizi hesablamaq üçün son modelimizi cari portfelə tətbiq edə bilərik.

Birinci məqaləni yekunlaşdırmaq üçün biz göstərdik ki, büzülmə üsullarından istifadə modelin dəqiqliyini artırır və bizə daha kiçik səhvlər verir. Ümumiyyətlə, ən köhnə modellərin inadkar istifadəsi tövsiyə edilmir. 2-ci hissədə əlavə araşdırmalar üçün LASSOizm və maşın öyrənməsi vasitəsilə xüsusiyyət seçimini daha dərindən araşdıracağıq.

İstinadlar:

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267–288.