Жіктеу модельдерін бағалау.

Теңгерімсіз деректерге арналған стратегиялар.

Модель диагностикасы

Басқа жіктеу әдістері сияқты, логистикалық регрессия модельдің жаңа деректерді қаншалықты дәл жіктейтіндігімен диагноз. Сызықтық регрессия сияқты, модельді диагностикалауға және жақсартуға мүмкіндік беретін бірнеше қосымша стандартты статистикалық құралдар бар. Бағалау коэффициенттерімен бірге R стандартты коэффициент қатесі (SE), z-балл және p-мәні туралы хабарлайды:

```
summary(logistic_model)

Call:
glm(formula = outcome ~ payment_inc_ratio + purpose_ + home_ +
    emp_len_ + borrower_score, family = "binomial", data = loan_data)

Deviance Residuals:

     Min        1Q    Median        3Q       Max
-2.71430  -1.06806  -0.04482   1.07446   2.11672

Coefficients:
                          Estimate  Std. Error z value Pr(>|z|)
(Intercept)               1.269822    0.051929  24.453  < 2e-16 ***
payment_inc_ratio         0.082443    0.002485  33.177  < 2e-16 ***
purpose_debt_consolidation 0.252164    0.027409   9.200  < 2e-16 ***
purpose_home_improvement  0.343674    0.045951   7.479 7.48e-14 ***
purpose_major_purchase    0.243728    0.053314   4.572 4.84e-06 ***
purpose_medical           0.675362    0.089803   7.520 5.46e-14 ***
purpose_other             0.592678    0.039109  15.154  < 2e-16 ***
purpose_small_business    1.212264    0.062457  19.410  < 2e-16 ***
home_OWN                  0.031320    0.037479   0.836 0.403
home_RENT                 0.168670    0.021041   8.016 1.09e-15 ***
emp_len_ < 1 Year         0.444892    0.053342   8.340  < 2e-16 ***
borrower_score           -4.638902    0.082433 -56.275  < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 64147  on 46271  degrees of freedom
Residual deviance: 58531  on 46260  degrees of freedom
AIC: 58555

Number of Fisher Scoring iterations: 4
```

P-мәнін түсіндіру регрессиядағыдай ескертулермен бірге жүреді және өзгермелі маңыздылықтың салыстырмалы көрсеткіші ретінде қарастырылуы статистикалық маңыздылықтың ресми өлшемі ретінде. RMS немесе R2 көрсеткіші екілік жауап беретін логистикалық регрессия моделімен байланысты емес. Оның орнына логистикалық регрессия моделі әдетте жіктеуге арналған жалпы метрикалық көрсеткіштермен бағаланады. Сызықтық регрессияға қатысты көптеген басқа ұғымдар логистикалық регрессияның параметрлік параметріне (және басқа OLM) ауыстырылады. Мысалы, қадамдық регрессияны қолдануға, өзара әрекеттесуді сипаттайтын теңдеу терминдерін сәйкестендіруге немесе сплайн терминдерін қосуға болады. Сол сұрақтар логистикалық регрессияға бұрмаланған және корреляцияланған айнымалыларды қолдануға қатысты. Жалпыланған аддитивті модельдердің кіші жарысын орындауға болады пакет арқылы

mgcv:
logistic_gam <- gam(outcome ~ s(payment_inc_ratio) + purpose_ + home_ + emp_len_ + s(borrower_score),
data=loan_data, family='binomial')

Логистикалық регрессия басқаша болатын салалардың бірі қалдықтарды талдауға қатысты. Регрессия сияқты (сурет. 4.9), жеке қалдықтарды есептеу тікелей орындалады:

terms <- predict(logistic_gam, type='terms')
partial_resid <- resid(logistic_model) + terms
df <- data.frame(payment_inc_ratio = loan_data[, 'payment_inc_ratio'], terms = terms[, 's(payment_inc_ratio)'],
partial_resid = partial_resid[, 's(payment_inc_ratio)'])
ggplot(df, aes(x=payment_inc_ratio, y=partial_resid, solid = FALSE)) + geom_point(shape=46, alpha=.4) +
geom_line(aes(x=payment_inc_ratio, y=terms),
color='red', alpha=.5, size=1.5) +
labs(y='Partial Residual')

Алынған график 5.4.-суретте көрсетілген. Сызықпен көрсетілген бағалау сәйкестігі нүктелік бұлттардың екі жиынтығы арасында өтеді. Жоғарғы бұлт 1 — жауапқа (қайтарылмаған несиелер) және төменгі бұлт 0 - жауапқа (шекті несиелер) сәйкес келеді. Бұл түр логистикалық регрессияның қалдықтарына тән, өйткені шығыс екілік болып табылады. Логистикалық регрессиядағы жеке қалдықтар регрессияға қарағанда аз мәнге ие болғанымен, олар сызықтық емес мінез-құлықты растау және өте ықпалды жазбаларды анықтау үшін әлі де пайдалы.

Логистикалық регрессияның негізгі идеялары:

- *Логистикалық регрессия сызықтық регрессияға ұқсас, тек жауап екілік айнымалы болып табылады.*
- *Коэффициент коэффициентінің логарифмі жауап айнымалысы ретінде әрекет ететін сызықтық формадағы модельді алу үшін бірнеше түрлендірулер қажет.*
- *Сызықтық модель орнатылғаннан кейін (итеративті процестің нәтижесінде) логарифмдік коэффициенттер ықтималдықтарға кері көрсетіледі.*
- *Логистикалық регрессия есептеу жылдамдығына байланысты танымал, себебі ол жаңа деректерді қайта есептемей-ақ бағалау үшін жедел қолдануға болатын модельді тудырады.*

Жіктеу модельдерін бағалау

Болжалды модельдеуде әр түрлі модельдердің көптігін сезіну, әрқайсысын кешіктірілген деректермен бақылау үлгісіне қолдану (сынақ немесе тексеру үлгісі деп те аталады) және олардың жұмыс қабілеттілігін диагностикалау әдеттегідей. Шын мәнінде, тәсіл олардың қайсысы ең дәл болжам жасайтынын байқауға келеді.

Негізгі терминдер:
Дәлдік (accuracy)
Дұрыс жіктелген жағдайлардың пайызы (немесе үлесі).
Сәйкессіздік матрицасы (confusion matrix)
Кестелік түрде көрсету (екілік жағдайда 2 = 2) олардың болжамды және нақты күйлері немесе жіктелу нәтижесі бойынша жазбалар саны.
Синонимдер: қате матрицасы, дәлсіздік матрицасы.
Сезімталдық (сезімталдық)
Дұрыс жіктелген бірліктердің пайызы (немесе үлесі).
Синоним: толықтығы.
Ерекшелік (ерекшелік)
Дұрыс жіктелген нөлдердің пайызы (немесе үлесі).
Дәлдік (дәлдік)
Болжалды бірліктердің пайызы (немесе үлесі), олар іс жүзінде нөлдер болып табылады.
ROC қисығы (ROC қисығы)
Сезімталдық графигі және ерекшелік.
Лифт (lift)
Әр түрлі ықтималдық шектерінде (салыстырмалы түрде сирек) бірліктерді анықтау кезінде модельдің тиімділік дәрежесін өлшейтін метрикалық көрсеткіш.

Жіктеу моделінің өнімділігін өлшеудің қарапайым әдісі - дұрыс болжамдардың үлесін есептеу. Жіктеу алгоритмдерінің көпшілігінде әр жағдайға "оның 1-ге тең болу

ықтималдығы" тағайындалады 3. Әдепкі бойынша, шешім қабылдау нүктесі немесе кесу әдетте 0,50 немесе 50% құрайды. Егер ықтималдық 0,5-тен жоғары болса, онда "1" санаты анықталады, әйтпесе - "0". Әдепкі балама кесу-бұл деректердегі бірліктердің басым ықтималдығы. Дәлдік (accuracy) — бұл жалпы қатенің өлшемі

$$\text{точность} = \frac{\Sigma_{\text{истинноположительный}} + \Sigma_{\text{истиннотрицательный}}}{\text{размер выборки}}.$$

Сәйкессіздік матрицасы

Жіктеудің метрикалық көрсеткіштері жүйесінің негізі - жауапсыздық матрицасы — жауап түрі бойынша санаттарға топтастырылған дұрыс және бұрыс аңыздардың санын көрсететін кесте. R-де сәйкессіздік матрицасын есептеуге арналған бірнеше бағдарламалық пакеттер бар, бірақ екілік жағдайда оны қолмен оңай есептеуге болады. Сәйкессіздік матрицасын көрсету үшін қайтарылмаған және өтелген несиелердің тең саны бар теңдестірілген деректер жиынтығында оқытылған logistic_game моделін қарастырыңыз (суретті қараңыз. 5.4). Қабылданған ережелерге сүйене отырып, 1 Y = мақсатты оқиғаға сәйкес келеді (мысалы, қайтарылмайды), ал 0 Y = теріс (немесе қалыпты) оқиғаға сәйкес келеді (мысалы, сөндірілген). Төмендегі код үзіндісі бүкіл (теңгерімсіз) жаттығу жиынтығына қолданылатын logistic_gam моделі үшін сәйкессіздік матрицасын есептейді:

```
pred <- predict(logistic_gam, newdata=train_set)
pred_y <- as.numeric(pred > 0)
true_y <- as.numeric(train_set$outcome=='default')
true_pos <- (true_y==1) & (pred_y==1)
 true_neg <- (true_y==0) & (pred_y==0)
false_pos <- (true_y==0) & (pred_y==1)
false_neg <- (true_y==1) & (pred_y==0)
conf_mat <- matrix(c(sum(true_pos),  sum(false_pos),
                     sum(false_neg), sum(true_neg)), 2, 2)
colnames(conf_mat) <- c('Yhat = 1', 'Yhat = 0')
rownames(conf_mat) <- c('Y = 1', 'Y = 0')
conf_mat
Yhat = 1 Yhat = 0
Y = 1 14635 8501
Y = 0 8236 14900
```

Болжалды нәтижелер-бағандар, ал шынайы нәтижелер — жолдар. Матрицаның диагональды элементтері дұрыс болжамдардың санын, диагональды емес элементтер қате болжамдардың санын көрсетеді. Мысалы, 6126 қайтарылмаған несиелер

қайтарылмағандар сияқты дұрыс болжанған және 17010 қайтарылмаған несиелер өтелгендер сияқты қате болжанған.

5.5-Суретте. Y екілік реакциясы мен әртүрлі метрикалық көрсеткіштер үшін сәйкессіздік матрицасы арасындағы байланысты көрсетеді (бөлімді қараңыз. Метрикалық көрсеткіштер туралы көбірек білу үшін осы тарауда "дәлдік, толықтық және ерекшелік"). Несие деректерінің мысалындағыдай, нақты жауап жолдар бойында, ал болжамды жауап бағандар бойында орналасқан. (Жолдар мен бағандардың инверттелген орналасуымен сәйкес келмейтін матрицаларды табуға болады.) Диагональды өрістер (сол жақ жоғарғы, оң жақ төменгі) ˆ Y болжамдары жауапты дұрыс болжаған кезде көрсетеді. Нақты айтылмаған маңызды метрикалық көрсеткіштердің бірі - жалған оң нәтижелердің үлесі (дәлдіктің айна бейнесі). Бірліктер сирек кездесетін болса, жалған оң нәтижелердің барлық болжамды оң нәтижелерге қатынасы жоғары болуы мүмкін, бұл қисынсыз жағдайға әкеледі, мұнда болжанған 1 0 болуы мүмкін. Бұл мәселе медициналық тексеруде (мысалы, маммограммалар) кеңінен қолданылатын диагностикалық сынақтардың қасіреті болып табылады: салыстырмалы түрде сирек кездесетіндіктен, сынақтардың оң нәтижелері сүт безі қатерлі ісігін білдірмейді. Бұл жұртшылықтың бағдарсыздығына әкеледі.

5.5. -Сурет. Екілік жауап пен әртүрлі метрикалық көрсеткіштер үшін сәйкессіздік матрицасы

Сирек кездесетін класс мәселесі.

Көптеген жағдайларда болжамды сыныптарда теңгерімсіздік бар, мұнда бір класс қалғандарына қарағанда әлдеқайда басым болады-мысалы, алаяқтарға қарсы жылқы сақтандыру талаптары немесе веб-сайттағы сатып алушыларға қарсы қарапайым келушілер. Сирек кездесетін класс (мысалы, алаяқтық шағымдар) - бұл әдетте үлкен қызығушылық тудыратын класс және әдетте 0 деп белгіленген үстемдікке қарағанда 1 деп белгіленеді. Әдеттегі сценарийде бірліктер маңызды жағдай болып табылады, өйткені оларды нөлдер ретінде дұрыс емес жіктеу нөлді бірлік ретінде дұрыс емес жіктеуге қарағанда қымбатырақ. Мысалы, алаяқтық сақтандыру талабын дұрыс анықтау мыңдаған долларды үнемдеуге мүмкіндік береді. Екінші жағынан, алаяқтық

емес шағымды дұрыс анықтау сізге мұқият талдаумен шағымды қолмен қарауға кететін шығындар мен күш - жігерді үнемдейді (егер сіз шағым "алаяқтық"деп белгіленген болса, дәл солай жасайсыз). Мұндай жағдайларда, егер сыныптар оңай бөлінбесе, жіктеудің ең дәл әдісі барлық жағдайларды 0 деп жіктейтін болуы мүмкін. Мысалы, Егер веб-дүкендегі қарапайым келушілердің тек 0,1% - ы сатып алуды аяқтаса, онда әрбір қарапайым сатып алушы сатып алусыз кетеді деп болжайтын модель 99,9% дәл болады. Дегенмен, ол пайдасыз болады. Оның орнына, біз жалпы дәлдігі аз, бірақ сатып алушыларды ажырата алатын модельге риза болар едік, тіпті егер ол кез - келген сатып алушыларды дұрыс жіктемесе де.

Дәлдік, толықтық және ерекшелік.

Таза дәлдіктен басқа метрикалық көрсеткіштер — неғұрлым нюансты сипаттағы көрсеткіштер-жіктеу модельдерін бағалауда кеңінен қолданылады. Олардың кейбіреулері статистикада ұзақ уақыт бойы, әсіресе биостатистикада қолданылады, онда олар диагностикалық сынақтардың күтілетін нәтижесін сипаттау үшін қолданылады. Дәлдік болжанған оң нәтиженің дәлдігін өлшейді (сурет. 5.5):

$$\text{Прецизионность} = \frac{\sum \text{ИП}}{\sum \text{ИП} + \sum \text{ЛП}},$$

мұндағы ЖК (ИП) — шынайы оң; LP(ЛП) -жалған оң. Толықтығы, сондай - ақ сезімталдық деп аталады, оң нәтижені болжаудағы модельдің Күшін өлшейді — ол дұрыс анықтайтын бірліктердің үлесі (сурет. 5.5). Сезімталдық термині биостатистика мен медициналық диагностикада жиі қолданылады, ал толықтығы Машиналық оқыту саласында көбірек қолданылады. Толықтығын анықтау келесі түрге ие.

$$\text{Полнота} = \frac{\sum \text{ИП}}{\sum \text{ИП} + \sum \text{ЛО}},$$

мұндағы ЖК (ИП)— шынайы оң; ЛО-жалған теріс.

Қолданылатын тағы бір метрикалық көрсеткіш-бұл модельдің теріс нәтижені болжау қабілетін өлшейтін ерекшелік

$$\text{Полнота} = \frac{\sum \text{ИО}}{\sum \text{ИО} + \sum \text{ЛО}},$$

мұндағы ИО-шын теріс; ЛО — жалған теріс.

```
# precision
conf_mat[1,1]/sum(conf_mat[,1])
 # recall
conf_mat[1,1]/sum(conf_mat[1,])
```

```
# specificity
conf_mat[2,2]/sum(conf_mat[2,])
```

ROC-қисық

Толықтығы мен ерекшелігі арасындаромаға келу бар екенін көруге болады. Бірліктердің көбірек санын қамту, әдетте, нөлдердің көп санын бірлік ретінде дұрыс анықтамауды білдіреді. Идеал жіктеуіш бірлік ретінде нөлдердің көп санын дұрыс анықтамай, бірліктерді сәйкестендірумен тамаша үйлеседі. Бұл компаны түсіретін метрикалық көрсеткіш әдетте ROC қисығы (receiver operating characteristics) деп аталатын "алушының жұмыс сипаттамалары" қисығы деп аталады. Кейінге қалдырудың у осі бойынша ROC қисығының графигін құру үшін толықтығы (сезімталдығы) х осінің ерекшелігіне сәйкес келеді.5 ROC қисығы жазбаны қалай жіктеу керектігін анықтай алу үшін сіз өзгерткен кесу шегіне қарай толықтық пен ерекшелік арасындағыроманы көрсетеді. Графикте сезімталдық (толықтық) у осінде көрсетіледі, ал х осі үшін таңбалаудың екі түрін табуға болады:

- -ерекшелік х осіне қолданылады, мұнда 1 сол жақта және 0 оң жақта;

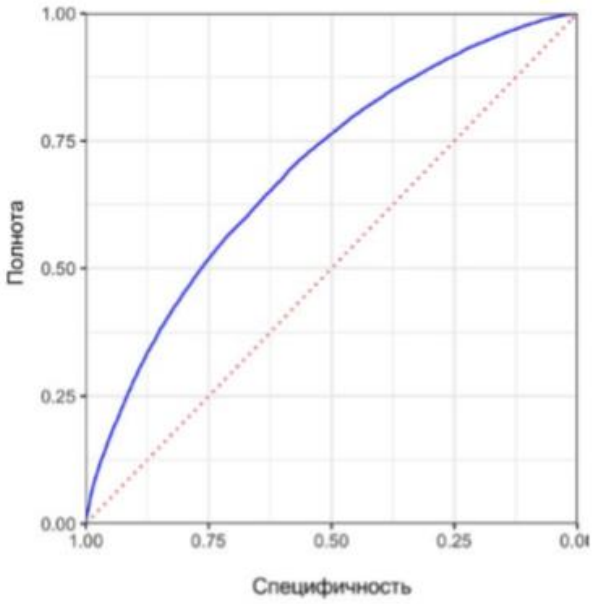
- -ерекшелік х осіне қолданылады, мұнда 0 сол жақта және 1 оң жақта.

Таңбалау формасына қарамастан қисық бірдей көрінеді. ROC қисығын есептеу процесі келесідей:
1. Жазбаларды болжамды ықтималдық бойынша сұрыптаңыз, 1-санатқа жататындығын көрсете отырып, ең ықтимал және ең аз ықтималдықпен аяқталады.
2. Сұрыпталған жазбалар негізінде жиынтық ерекшелік пен толықтықты есептеңіз.

R-дегі ROC қисығын есептеу жеткілікті түрде түзу орындалады. Төмендегі код үзіндісі несие деректері үшін ROC есептейді:

```
idx <- order(-pred)
recall <- cumsum(true_y[idx]==1)/sum(true_y==1)
specificity <- (sum(true_y==0) - cumsum(true_y[idx]==0))/sum(true_y==0)
roc_df <- data.frame(recall = recall, specificity = specificity)
ggplot(roc_df, aes(x=specificity, y=recall)) +
geom_line(color='blue') +
scale_x_reverse(expand=c(0, 0)) +
scale_y_continuous(expand=c(0, 0)) +
geom_line(data=data.frame(x=(0:100)/100), aes(x=x, y=1-x),
linetype='dotted', color='red')
```

Нәтиже суретте көрсетілген. 5.6. Нүктелі диагональды сызық кездейсоқ мүмкіндіктен жақсы емес классификаторға сәйкес келеді. Өте тиімді жіктеуіш (немесе медициналық жағдайларда өте тиімді ди - агностикалық тест) жоғарғы сол жақ бұрышқа басылған ROC қисығына ие болады-ол көптеген нөлдерді бірлік ретінде қате жіктеместен көптеген бірліктерді дұрыс анықтайды. Егер бұл модель үшін бізге кем дегенде 50% спецификациясы бар классификатор қажет болса, онда толықтығы шамамен 75% құрайды.



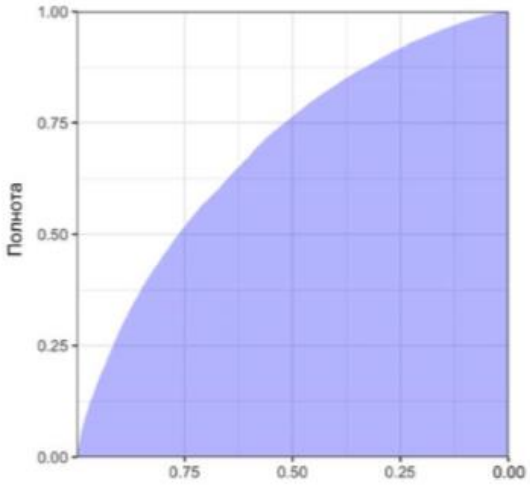
5.6. - Сурет. ROC - несие деректерінің қисығы

AUC метрикалық көрсеткіші

ROC қисығы құнды графикалық құрал болып табылады, бірақ ол жіктеуіштің өнімділігінің жалғыз өлшемін білдірмейді. ROC қисығын AUC (area under the ROC curve) метрикалық көрсеткішін жасау үшін пайдалануға болады. AUC метрикалық көрсеткіші - бұл ROC қисығының астындағы жалпы аудан. AUC мәні неғұрлым үлкен болса, жіктеуіш соғұрлым тиімді болады. 1-ге тең AUC идеалды классификатор туралы айтады: ол барлық 1 - ді дұрыс анықтайды және кез-келген 0-ді 1-ге теңестірмейді. Мүлдем тиімсіз жіктеуіш-диагональды сызық-0,5-ке тең AUC болады. Суретте. 5.7 несие моделі үшін ROC қисығының астындағы аймақ көрсетілген. AUC мәнін сандық интеграция арқылы есептеуге болады:

*sum(roc_df$recall[-1] **

diff(1-roc_df$specificity)) [1] 0.5924072



5.7.- Сурет. Несие деректері үшін ROC қисығы (AUC) астындағы аймақ

Лифт

AUC мәнін метрикалық көрсеткіш ретінде пайдалану қарапайым дәлдікпен салыстырғанда жақсарту болып табылады, өйткені бұл көрсеткіш жіктеуіштің өзара дәлдік пен маңызды бірліктерді анықтау қажеттілігі арасындағыроманы қаншалықты жақсы өңдейтінін анықтай алады. Алайда, бұл сирек кездесетін жағдай мәселесін толығымен шешпейді, мұнда барлық жазбалардың 0 - ге сәйкестендірілуіне жол бермеу үшін модельдің кеуекті ықтималдығын 0,5-тен төмендету қажет. Мұндай жағдайларда жазбаны 1-ге жіктеу үшін 0,4; 0,3 немесе одан төмен ықтималдығы жеткілікті болуы мүмкін. Нәтижесінде біз олардың маңыздылығын көрсете отырып, бірліктерді шамадан тыс анықтайтын нәрсеге келеміз.

Бұл кесу шегін өзгерту бірліктерді қамту мүмкіндігін арттырады (нөлдердің көп санын бірлік ретінде дұрыс жіктемеу арқылы). Бірақ кесудің оңтайлы шегі қандай?

Лифт ұғымы бұл сұрақтың жауабын кейінге қалдыруға мүмкіндік береді. Оның орнына жазбалар олардың болжамды ықтималдығы бойынша Бірлік санатына жатады. Айталық, алгоритм бірлік ретінде анықталған жоғарғы 10% - да сынып тек соқыр таңдалған критериймен салыстырғанда қаншалықты жақсы жұмыс істеді? Егер сіз осы жоғарғы децилде 0,1% жауаптың орнына 0,3% жауап ала алсаңыз, онда алгоритмде лифт бар (немесе пайда, ағылшын тілінен). gains), жоғарғы децилде 3-ке тең. Лифт графигі (өсу графигі) мұны деректер ауқымында анықтайды. Мұндай графикті деректер ауқымында арнайы немесе үздіксіз құруға болады.

Лифт графигін есептеу үшін алдымен y осіндегі толықтығын және X осіндегі жазбалардың жалпы санын көрсететін жиынтық өсу графигі құрылады. Спецификалық өсу кестелері-бұл интернет-коммерция пайда болғанға дейінгі кезеңге жататын болжамды модельдеудегі ең көне әдістердің бірі. Олар әсіресе тауарларды пошта арқылы сататын кәсіпқойлар арасында танымал болды. Сатудың бұл түрі жарнаманың қымбат әдісі болып табылады, егер ол оқылмайтын түрде қолданылса және жарнама берушілер болжамды модельдерді (алғашқы жылдары өте қарапайым) ықтимал төлем перспективасымен әлеуетті клиенттерді анықтау үшін қолданды.

***Жіктеу модельдерін бағалаудың негізгі идеялары**: •*
Дәлдік (дұрыс болжанған сәйкестендірулердің пайызы) модельді бағалаудың алғашқы қадамы ретінде қарастырылуы керек. •
Басқа метрикалық көрсеткіштер (толықтығы, ерекшелігі, дәлдігі) тиімділіктің негұрлым нақты сипаттамаларына бағытталған (мысалы, толықтығы модельдің бірліктерді дұрыс сәйкестендіруде қаншалықты жақсы жұмыс істейтінін өлшейді). •
AUC (ROC қисығының астындағы аймақ) - модельдің бірліктерді нөлдерден ажырату қабілетінің жалпы қабылданған метрикалық көрсеткіші. •
Сол сияқты, лифт модельдің бірліктерді анықтауда қаншалықты тиімді екенін өлшейді және ол көбінесе ықтимал бірліктерден бастап арнайы есептеледі.

Теңгерімсіз деректерге қатысты стратегиялар

Алдыңғы бөлімде қарапайым дәлдіктен асып түсетін және теңгерімсіз деректерге жарамды метрикалық көрсеткіштер бойынша жіктеу модельдерін бағалау қарастырылды - мақсатты нәтиже (веб — сайтта шомылу, сақтандыру алаяқтықтары және т.б.) сирек кездесетін деректер. Бұл бөлімде біз теңгерімсіз деректермен болжамды модельдеудің өнімділігін жақсартатын қосымша стратегияларға жүгінеміз.

Негізгі терминдер
Төмен үлгі (undersample) классификация моделінде басым класы бар жазбаларды азырақ пайдаланыңыз.
Күшейту үлгісі (oversample) классификация моделінде сирек кездесетін сыныптармен көбірек жазбаларды қолданыңыз, қажет болған жағдайда жүктеу көмегіне жүгініңіз.
Артық салмақтың жоғарылауы немесе төмендеуі (up weight or down weight) үлгідегі сирек (немесе басым) класс - суға үлкен (немесе кіші) салмақ тағайындау.
Деректерді құру (data generation) жүктеуге ұқсас Процедура, тек әрбір жаңа жүктеу жазбасы өз көзінен сәл өзгеше.
Z-балл (z-score) стандарттаудан кейін алынған мән.
K
Жақын көршілерді есептеу кезінде ескерілетін көршілердің саны.

Төмен таңдау

Егер деректер жеткілікті болса, несие деректері сияқты, шешімдердің бірі басым сыныпты іріктеуді төмендетуде ат - дан тұрады, нәтижесінде модельдеу деректері нөлдер мен бірліктер арасында теңдестірілген болады. Төменгі таңдаудың негізгі идеясы-доминантты сыныпқа арналған деректерде көптеген артық жазбалар бар. Кішірек, теңдестірілген деректер жиынтығымен жұмыс істеу модельдің тиімділігінің

артықшылықтарына әкеледі және деректерді дайындауды, сондай - ақ эксперименттік модельдерді зерттеу мен сынауды жеңілдетеді. Қанша деректер жеткілікті болады? Бұл қосымшаға байланысты, бірақ жалпы алғанда, онша басым емес сынып үшін ондаған мың жазбалардың болуы жеткілікті болады. Бірліктер мен нөлдер неғұрлым оңай анықталса, соғұрлым аз мәліметтер қажет. Бөлімде талданған несие деректері. "Логистикалық регрессия" осы тараудың басында теңдестірілген жаттығулар жиынтығына негізделген: жартысы несиеге төленді, ал екінші жартысы қайтарылмады. Болжалды мәндер ұқсас болды, ықтималдықтардың жартысы 0,5 — тен аз, ал жартысы 0,5-тен жоғары болды. Толық деректер жиынтығында несиелердің шамамен 5% - ы ғана қайтарылмады:

```
mean(loan_all_data$outcome == 'default')
[1] 0.05024048
```

Модельді жаттықтыру үшін толық деректер жиынтығын пайдалансаңыз не болады?

```
full_model <- glm(outcome ~ payment_inc_ratio + purpose_ +
home_ + emp_len_+ dti + revol_bal + revol_util,
data=train_set, family='binomial')
pred <- predict(full_model)
mean(pred > 0)
[1] 0.00386009
```

Несиелердің тек 0,39% - ы қайтарылмайтын болады немесе күтілетін санның 1/12-ден аз болады деп болжанған. Өтелген несиелер саны бойынша басылады қайтарылмаған несиелер, өйткені модель барлық оди - наково деректерін қолдана отырып дайындалған. Егер сіз бұл туралы интуитивті деңгейде ойласаңыз, онда мұндай қайтарылмайтын несиелердің болуы, болжамды деректердегі сөзсіз вариациямен бірге, тіпті қайтарылмайтын несие үшін де модель бірнеше қайтарылмайтын несиелерді табуы мүмкін дегенді білдіреді.олар кездейсоқ түрде ұқсас болады. Теңдестірілген іріктеу қолданылған кезде несиелердің шамамен 50% қайтарылмайтын деп болжанған.

Таңдауды жоғарылату және жоғарылату/төмендету

Жазбаларды іріктеуді төмендету әдісі туралы бір сын-бұл деректерді тастайды және қолда бар барлық ақпаратты пайдаланбайды. Егер сізде салыстырмалы түрде шағын деректер жиынтығы болса және сирек болса сынып бірнеше жүз немесе бірнеше мың жазбалардан тұрады, содан кейін басым сыныптың пони таңдауы пайдалы ақпаратты тастау қаупі бар. Бұл жағдайда доминантты жағдайды төмендетудің орнына, қайтарылатын қосымша жолдарды алу (жүктеу) арқылы сирек кездесетін сыныпты жоғарылату керек. Деректерді өлшеу арқылы ұқсас әсерге қол жеткізуге болады. Көптеген жіктеу алгоритмдері деректерді жоғарылатуға/төмендетуге мүмкіндік беретін салмақ дәлелін қабылдайды. Мысалы, glm-ге салмақ аргументі арқылы несие деректеріне салмақ векторын қолдану:

```
wt <- ifelse(loan_all_data$outcome=='default',
1/mean(loan_all_data$outcome == 'default'), 1)
full_model <- glm(outcome ~ payment_inc_ratio + purpose_ +
home_ + emp_len_+ dti + revol_bal + revol_util,
data=loan_all_data, weight=wt, family='binomial')
pred <- predict(full_model)
mean(pred > 0)
 [1] 0.4344177
```

Қайтарылмайтын несиелер үшін салмақтар 1/p-ге орнатылады, мұндағы p-қайтарылмау ықтималдығы. Қайтарылмайтын соттардың салмағы 1. Қайтарылмайтын несиелер мен қайтарылмайтын несиелер үшін салмақ сомалары шамамен баламалы. Болжалды мәндердің орташа мәні қазір 0,39% орнына 43% құрайды. Айта кету керек, артық салмақ сирек кездесетін сыныпты жоғарылату үшін де, басым сыныпты төмендету үшін де балама ұсынады.

Деректерді құру

Деректерді генерациялау-бут-страпинг арқылы жазбаларды іріктеуді жоғарылату нұсқасы (бөлімді қараңыз. "Төмен таңдау" осы тараудың басында) жаңа жазбалар жасау үшін бар жазбаларды қайта қараумен. Бұл идеяның негізінде жатқан логика мынада: біз тек жағдайлардың шекті жиынтығын байқайтындықтан, алгоритмде класс - сификацияның "ережелерін" құру үшін бай ақпарат жиынтығы жоқ. Қолданыстағы жазбаларға ұқсас, бірақ бірдей емес жаңа жазбалар жасау арқылы алгоритм ұялшақ ережелер жиынтығын үйренуге мүмкіндік алады. Бұл ұғым ансамбльдік статистикалық сәнге ұқсас, атап айтқанда бэггинг және бустинг (6 тарауды қараңыз). Бұл идея smote алгоритмін жариялаумен қарқын алды, оның атауы синтетикалық азшылықты (synthetic minority oversampling technique) іріктеуді арттыру әдістемесі ретінде түсініледі. SMOTE алгоритмі таңдаудың жоғарылауына ұшыраған жазбаға ұқсас жазбаны табады (бөлімді қараңыз. 6 - тараудың" k жақын көршілері"), және синтетикалық жазбаны жасайды, ол бастапқы жазбаның және көрші жазбаның кездейсоқ өлшенген орташа мәні болып табылады, Мұнда Салмақ әр болжаушы үшін бөлек жасалады. Іріктеуді жоғарылату арқылы алынған синтетикалық жазбалардың саны деректер жиынтығын нәтижелер кластарына қатысты шамамен тепе-теңдікке келтіру үшін қажет болатын іріктеу коэффициентіне байланысты. R-де бірнеше SMOTE іске асырулары бар. Теңгерімсіз деректерді өңдеуге арналған ең жан-жақты бағдарламалық пакет-теңгерімсіз. Ол ең жақсы әдісті таңдау үшін жарыс алгоритмін ("жарыс") қоса алғанда, әртүрлі мамандандырылған әдістерді ұсынады. Алайда, SMOTE алгоритмі өте қарапайым және оны knn пакетін пайдаланып R-ге тікелей енгізуге болады.

Құнға бағытталған жіктеу

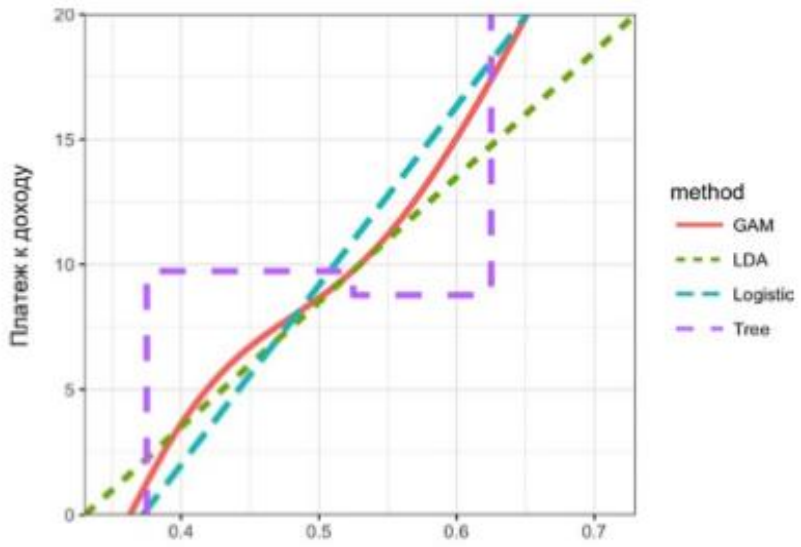
Іс жүзінде дәлдік көрсеткіштері мен AUC — бұл жіктеу ережесін таңдаудың жеңілдетілген әдістері. Көбінесе бағалау құны жалған теріс нәтижеге қарсы жалған нәтижеге тағайындалады және бұл мәндерді бірліктер мен нөлдерді жіктеу кезінде ең жақсы кесу шегін анықтау үшін қосқан жөн. Мысалы, жаңа несиені қайтармаудың болжамды құны C, ал өтелген несиеден күтілетін кіріс R деп есептейік.:

$$\text{ожидаемый доход} = P(Y=0)\cdot R + P(Y=1)\cdot C.$$

Несиені қайтарылмайтын немесе өтелмеген деп белгілеудің немесе қайтарылмау ықтималдығын анықтаудың орнына, несиенің оң күтілетін кірісі бар - жоғын анықтаған дұрыс. Қайтарылмаудың болжамды ықтималдығы аралық қадам болып табылады және бизнестегі түпкілікті жоспарлы метрикалық көрсеткіш болып табылатын күтілетін пайданы анықтау үшін несиенің қорытынды құнымен біріктірілуі керек. Мысалы, кішігірім несиені қайтарымсыз болжау ықтималдығы сәл жоғары үлкен несиенің пайдасына үнсіздікпен айналып өтуге болады.

Болжамдарды зерттеу

AAC сияқты бір метрикалық көрсеткіш модельдің белгілі бір жағдайға сәйкестігінің барлық аспектілерін қамти алмайды. Суретте. 5.8 тек екі болжамды айнымалыны қолданатын кеме деректеріне сәйкес келетін төрт түрлі модельге арналған шешім ережелері көрсетілген: borrower_score және payment_inc_ratio. Модельдер: сызықтық дискриминантты талдау (LDA), логистикалық сызықтық регрессия, жалпыланған аддитивті модель (GAM) көмегімен орнатылған логистикалық регрессия және ағаш моделі (бөлімді қараңыз. 6 - тараудың" ескі модельдері"). Сызықтардың жоғарғы сол жағындағы аймақ болжамды қайтаруға сәйкес келеді. Бұл жағдайда LDA және логистикалық сызықтық регрессия бірдей нәтиже береді. Ағаш моделі ең аз тұрақты ережені шығарады: іс жүзінде, бағалаушының балының өсуі болжамды "өтелгеннен" "қайтарылмайтын"бағытқа ауыстыратын жағдайлар бар! Сонымен, GAM негізіндегі логистикалық регрессияға сәйкестік ағаш пен сызықтық модельдер арасындағы компаны білдіреді.



5.8. -сурет. Төрт түрлі әдіс үшін жіктеу ережелерін салыстыру

Болжау ережелерін жоғары өлшемдерде немесе GAM және ағаш үлгісі жағдайында елестету, тіпті мұндай ережелер үшін аймақтарды құру өте қиын. Қалай болғанда да, болжамды мәндерді барлау талдауы әрқашан негізделген.

Теңгерімсіз деректер стратегияларының негізгі идеялары:

- *Өте теңгерімсіз деректер (яғни қызығушылық нәтижелері, бірліктер сирек кездесетін жерде) жіктеу алгоритмдеріне қиындық тудырады.*
- *Стратегиялардың бірі - жаттығу деректерін көптеген жағдайды төмендету арқылы теңестіру (немесе сирек кездесетін жағдайды жоғарылату).*
- *Егер барлық бірліктерді пайдалану сізге әлі де тым аз бірлік берсе, онда сирек кездесетін жағдайларды жүктеуді қолдануға болады немесе бар сирек жағдайларға ұқсас синтетикалық деректерді жасау үшін SMOTE алгоритмін қолдануға болады.*
- *Теңгерімсіз деректер әдетте бір сыныпты (бірліктерді) дұрыс сәйкестендірудің құны жоғары екенін және ди - агностикалық метрикалық көрсеткішке құндылық коэффициентін енгізу керек екенін көрсетеді.*