Predikce hospitalizační mortality u akutního infarktu myokardu

Václav Kratochvíl, Hynek Kružík, Petr Tůma, Jiří Vomlel a Petr Somol

Anotace

Předmětem práce je standardizace výsledkového ukazatele „Nemocniční mortalita při akutním infarktu myokardu“ s využitím zjištěných závislostí mezi dílčími rizikovými faktory pacienta a úmrtím pacienta.

Klíčová slova

Standardizace rizika, hodnocení nemocniční péče, logistická regrese, strojové učení, dolování z dat.

1. Úvod

Výsledky lékařské péče závisejí nejen na vhodném výběru a správném provedení léčebných postupů, ale také na výchozím stavu pacienta. Vyhodnocení vstupního stavu pacienta je v praxi použitelné pro dva typy úloh:

1. pro odhad prognózy konkrétního pacienta a rozhodnutí a adekvátním postupu spočívajícím například v použití určitých intervencí, léků nebo překladu pacienta na pracoviště vyššího typu.
2. pro zpětné statistické hodnocení výsledků péče obvykle za použití ustálených ukazatelů kvality.

Pro prvý případ použití se často požívá pojem *Stratifikace rizik*, zatímco pro druhý způsob je obecně přijímán pojem *Standardizace rizika*. Je ovšem třeba si uvědomit, že koncepčně i metodicky musí být oba postupy plně konzistentní. Standardizace rizika výsledkových ukazatelů kvality je v podstatě založen na stratifikaci rizik a vychází z empirických i exaktních poznatků o vlivu jednotlivých rizikových faktorů u individuálního pacienta na globálně vyjádřený výsledek péče u skupiny pacientů.

V praktickém provedení se individuální Stratifikace rizika liší od Standardizace rizika hlavně tím, že vyhodnocení individuálního rizika se provádí v reálném čase, zatímco Standardizace rizika retrospektivně. V individuálním případě jsou proto k dispozici všechny údaje z klinické dokumentace (anamnestické údaje, fyziologické hodnoty, nálezy pomocných vyšetření) a pokud některý v dokumentaci není, je možné jej zjistit (doplnit anamnézu nebo vyšetření). Retrospektivní skupinové hodnocení kvality výsledku podléhá při Standardizace rizika řadě omezení. Nejen že není možné doplnit chybějící údaje, ale je nutné pracovat s tím, co nabízí dostupný zdroj informací, což je obvykle určitý datový soubor, velmi často definovaný pro nějaké rutinní použití (jiné než měření kvality). V praxi je tedy často jediným zdrojem administrativní datový soubor, používaný pro vykazování péče zdravotním pojišťovnám, nebo vykazování pro „státní statistiku“ (ÚZIS). Modely, založené na těchto „rutinně“ sbíraných datech, se často nazývají „administrativními“ modely.

V této práci se zabýváme vyhodnocením rizik v metodě Standardizace rizika. Vybrali jsme si ukazatel „Nemocniční mortalita při akutním infarktu myokardu“, protože:

- Tento ukazatel je spolehlivě zjistitelný z běžných dat.
- Léčba akutního infarktu myokardu je klinickou oblastí, která je celospolečensky významná a má i nezanedbatelný ekonomický dopad.
- Predikce mortality na základě vstupních nálezů je předmětem řady odborných prací.

V běžné praxi se v českém prostředí ukazatele predikce výsledku léčby běžně nepoužívají. Stejně tak nebyl v českém prostředí ověřován žádný model standardizace ukazatele „Nemocniční mortalita při akutním infarktu myokardu“.

2. Standardizace rizika

Standardizace je statistická metoda úpravy výsledku měření, jejímž cílem je lepší možnost porovnání a interpretace. Standardizují se především ukazatele výsledkové, v některých případech i ukazatele procesní. Kvalitu poskytované péče lze měřit několika typy výsledkových ukazatelů, mezi něž patří například ukazatele mortality, relativní počty opakovaných hospitalizací a relativní počty komplikací. V této části ukážeme důvody a postup standardizace na ukazateli typu mortalita; pro ostatní výsledkové ukazatele platí stejné standardizační principy.

Přestože mezi sebou porovnáváme zdravotnická zařízení pouze v rámci jednoho klinicky relativně specifického ukazatele (např. nemocniční mortalita u akutního infarktu myokardu), není takovéto srovnání samo o sobě (bez dalších kroků) objektivní. Rozdíl v hodnotě mortality je nejen důsledkem rozdílné aplikace správných léčebných postupů, ale také důsledkem různé skladby přijímaných a léčených pacientů. Jinými slovy nemocnice přijímající komplikovanější nebo rizikovější případy budou mít patrně vyšší mortalitu, aniž by to znamenalo, že poskytují méně kvalitní péči.

Rizikové faktory pacienta, které nemusejí mít vždy pouze klinickou povahu, jsou důvodem zkreslení ukazatele. Zkreslení ukazatele je způsobeno nerovnoměrnou distribucí rizikových faktorů mezi jednotlivé poskytovatele. Zkreslení ukazatele je nutné odlišovat od nepřesnosti ukazatele, která je způsobená nesprávným sběrem dat, nesprávným způsobem zjišťování faktů apod. Zkreslení je vždy dáno faktory reálného světa, jejich přítomnost není výsledkem poruchy měření, sběru dat apod. Smyslem a cílem standardizace ukazatele je odstranění zkreslení.

Pokud je homogenita měřené skupiny (z hlediska výskytu rizikových faktorů) nedostatečná, existují dva základní postupy, jak hodnotu měření zlepšit z hlediska korektní interpretace:

- *stratifikace* je rozklad na podskupiny (například podle věkových skupin) a provedení samostatných měření pro ně. Toto je ovšem nevýhodné jak z hlediska statistiky (podskupiny budou mít často malé počty pacientů), tak z hlediska následné interpretace (různé podskupiny mohou mít různé komparativní výsledky a nemusí být tedy jasné, jak si poskytovatel vlastně stojí v celkové kvalitě pro danou klinickou oblast).
- *standardizace* je matematická operace, která odstraní vliv rizikových faktorů tak, aby zůstal jeden (syntetický) výsledek.

Standardizace tedy odstraňuje zkreslení výsledkového ukazatele, které je způsobené nerovnoměrnou distribucí rizikových faktorů mezi jednotlivé poskytovatele. Cílem standardizace je vyloučit matematickým postupem vliv těch rizikových faktorů na straně pacienta, které existovaly již v době příjmu do nemocnice (nebo jinak stanoveného začátku posuzované epizody) a které negativně ovlivňují hodnoty výsledkových ukazatelů.

3. Výběr rizikových faktorů použitých při standardizaci

Požadavky kladené na výběr rizikových faktorů použitelných v procesu standardizace ukazatelů jsou následující:

- Musí existovat statisticky doložitelná závislost mezi rizikovým faktorem a hodnotou výsledkového ukazatele - např. pravděpodobnost úmrtí na infarkt myokardu souvisí s hodnotou krevního tlaku při příjmu. Matematický model by měl zohlednit, pokud určité kombinace faktorů mají větší „váhu“, než odpovídá „součtu“ vlivů jednotlivých faktorů.
- Zdravotnická zařízení se musí lišit skladbou přijímaných pacientů z pohledu rizikového faktoru. V opačném případě není zapotřebí standardizaci provádět, třebaže existuje silná závislost mezi ukazatelem a rizikovým faktorem – všechna zdravotnická zařízení jsou totiž „znevýhodněna“ stejným způsobem. V praxi je tento požadavek většinou splněn, neboť zdravotnická zařízení se mezi sebou obvykle liší distribucí rizikového faktoru.
- Rizikový faktor musí jednoznačně odrážet stav pacienta již při příjmu pacienta do zdravotnického zařízení a nesmí být výsledkem samotného léčebného procesu.
- O rizikovém faktoru musí existovat spolehlivé záznamy. Pokud bude systém založen na administrativních datech (např. dávkách pro zdravotní pojišťovny), musí být rizikový faktor standardně z těchto dat dostupný. Toto je bohužel velmi limitující omezení. I pokud bychom věděli, které faktory nejlépe odrážejí stav pacienta a jeho riziko pro nepříznivý výsledek v momentě příjmu, nebude snadné je získat, protože většinou nejsou v administrativních datech vůbec, nebo jsou tam v nespolehlivé kvalitě.

Odstranění zkreslení není nikdy úplné. I po aplikaci standardizačních postupů zůstává zbytkové zkreslení. Zbytkové zkreslení výsledku může být způsobeno například tím, že některé rizikové faktory nejsou ještě známy, nebo nejsou podchyceny v datech.

Rizikové faktory, které mají vliv na pravděpodobnost úspěchu léčby, závisí:

- na neklinických charakteristikách (sociální statut, etnické a kulturní odlišnosti),
- na demografických charakteristikách (věk, pohlaví) a
- na klinických charakteristikách.

4. Jak se ukazatele standardizují?

Při standardizaci je třeba nejprve v první fázi nalézt a vyjádřit vztah mezi výsledkovým ukazatelem a rizikovým faktorem. Dejme tomu, že rizikovým faktorem je věk a výsledkovým ukazatelem je *Nemocniční mortalita u akutního infarktu myokardu*. Na základě dat z celého souboru zdravotnických zařízení (standardní populace) je pak zapotřebí za pomoci statistických metod vyjádřit vztah mezi věkem a pravděpodobností úmrtí na infarkt.

V další fázi standardizace je pro každou nemocnici vypočten tzv. srovnávací index (SI), který je definován jako podíl 2 veličin: skutečného počtu úmrtí a predikovaného (očekávaného) počtu úmrtí:

SI = skutečný počet úmrtí / predikovaný (očekávaný) počet úmrtí.

Predikovaný (očekávaný) počet úmrtí získáme tak, že pro každého pacienta příslušné nemocnice vypočteme (v závislosti na hodnotě rizikového faktoru) pravděpodobnost úmrtí a tyto pravděpodobnosti sečteme přes všechny pacienty nemocnice. Interpretace této veličiny je následující: jedná se o počet úmrtí, které bychom očekávali, pokud by v dané nemocnici platili stejné úmrtnostní zákony jako v celé populaci nemocnic. Pokud tento počet porovnáme se skutečným počtem zemřelých, dostáváme odpověď na otázku, zda v nemocnici bylo více nebo méně úmrtí, než

by bylo možné očekávat na základě distribuce rizikového faktoru u pacientů. Samotný srovnávací index je bezrozměrné číslo, které udává relativní pozici nemocnice ve srovnání s průměrem: hodnota indexu větší než jedna značí nadprůměrnou úmrtnost v nemocnici, hodnota indexu menší než jedna značí naopak podprůměrnou úmrtnost. Abychom se dostali zpět do úrovně původních hodnot úmrtnosti, je třeba tento index vynásobit hodnotou obecné úmrtnosti, tj. průměrnou úmrtností za všechny nemocnice:

Standardizovaná úmrtnost = obecná úmrtnost * SI.

Při standardizaci je tedy výsledkem zjištění, jak by vypadal ukazatel, kdyby u daného poskytovatele zdravotnických služeb bylo stejné zastoupení rizikových faktorů, jako v celém souboru a kdyby platily stejné souvislosti mezi rizikovými faktory a ukazatelem, jako v celém souboru.

5. Standardizace metodou logistické regrese

Pro experimenty jsme použili datový soubor z jedné anonymní nemocnice v České republice, který obsahoval záznamy o pacientech s přijímací diagnózou akutního infarktu myokardu (I210 až I214). Po vyřazení pacientů, kteří byli přeloženi do jiné nemocnice, datový soubor obsahoval celkem 486 pacientů. U 335 pacientů chyběla hodnota některého z vybraných rizikových faktorů. Pacienty s chybějícími záznamy jsme nevyřadili, ale pro zachování maxima využitelné informace jsme použili metodu pro vkládání chybějících údajů zmíněnou později v této kapitole.

Prvním a nejsložitějším krokem při standardizaci vybraného ukazatele je nalézt relevantní rizikové faktory a matematicky charakterizovat jejich vliv na vybraný ukazatel. Obvykle (viz například [1]) se vztah mezi rizikovými faktory a vybraným ukazatelem vyjadřuje pomocí logistické regrese. Nechť $P(Y=1)$ značí pravděpodobnost, že veličina Y dosáhne hodnoty 1. V našem případě to bude pravděpodobnost, že pacient do 30 dnů po přijetí do nemocnice zemře. Model logistické regrese definuje vztah mezi závislou veličinou Y a vektorem rizikových faktorů $\boldsymbol{X}$ nabývajícího vektoru hodnot $\boldsymbol{x}$. Vztah je definován pomocí logistické funkce jako

$$P(Y=1) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x})},$$

kde $\boldsymbol{\beta}$ je vektor parametrů, které je třeba při učení modelu nalézt a $\boldsymbol{\beta}'$ značí jeho transpozici. Vektor hodnot $\boldsymbol{x}$ má obvykle tvar $(1, \boldsymbol{z})$ a první složka vektoru $\boldsymbol{\beta}$ označovaná β_0 je takzvaný absolutní člen (intercept).

Při výběru rizikových atributů jsme vycházeli ze záznamů uchovaných v nemocničním informačním systému ve strukturované podobě. Měli jsme k dispozici výsledky laboratorních testů, hlavní a vedlejší diagnózy i roční historii diagnóz jednotlivých pacientů v dané nemocnici, celkem 637 položek. Samozřejmě, že pro každého jednotlivého pacienta nebyla většina laboratorních testů provedena a většina diagnóz byla negativní (kódována hodnotou 0). Prvním krokem byl výběr kandidátů na rizikové faktory. K tomu jsme použili metodu založenou na *informačním zisku*, který je pro každý rizikový faktor X a závislou veličinu Y definován

$$\mathrm{I}(\mathrm{Y}, X) = \mathrm{H}(X) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(X, \mathrm{Y}),$$

kde $\mathrm{H}(X)$ je entropie veličiny X definovaná jako

$$\mathrm{H}(X) = -\sum_{x} \mathrm{P}(\mathrm{X} = \mathrm{x}) \log \mathrm{P}(\mathrm{X} = \mathrm{x})$$

a $\mathrm{H}(X, Y)$ je sdružená entropie veličin X a Y definovaná obdobně jako

$$\mathrm{H}(X, Y) = -\sum_{x,y} \mathrm{P}(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) \log \mathrm{P}(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) ,$$

kde sčítáme přes všechny kombinace stavů x, y veličin X, Y. Logaritmus používáme binární. Čím vyšší je informační zisk, tím více informace nám veličina X přináší o veličině Y. Výsledky laboratorních testů vstupovaly do analýzy svými absolutními hodnotami (nikoliv relativními hodnotami vzhledem k normálu pro daný věk). Hodnoty spojitých veličin byly pro účely výpočtu informačního zisku diskretizovány do deseti košů. Do dalšího zpracování jsme zařadili pouze ty veličiny, jejichž informační zisk byl větší než $0{,}01$. Vybrané veličiny jsou uvedeny v Tab. 1.

Kód	Popis	Informační zisk
S.Urea	Močovina v séru	0.11674202
S.kreatinin	Kreatinin v séru	0.09532763
B.leukocyty	Leukocyty v plné krvi	0.06820710
I48	Fibrilace a flutter síní	0.02425088
O.E78	Poruchy metabolismu lipoproteinů a jiné lipidémie	0.02318073
O.I20	Angina pectoris	0.02044021
O.I48	Fibrilace a flutter síní	0.01997538
I73	Jiné nemoci periferních cév	0.01971532
O.I27	Jiné kardiopulmonální nemoci	0.01971532
O.I73	Jiné nemoci periferních cév	0.01971532
Vek	Věk pacienta	0.01926587
O.I46	Srdeční zástava	0.01851840
K92	Jiné nemoci trávicí soustavy	0.01758336
O.I210	Akutní transmurální infarkt myokardu přední stěny	0.01651995
I74	Tepenný vmetek - arteriální embolie - a trombóza	0.01576957
I42	Kardiomyopatie	0.01474711
O.I42	Kardiomyopatie	0.01474711
O.I10	Esenciální (primární) hypertenze	0.01471863
O.I211	Akutní transmurální infarkt myokardu spodní (dolní) stěny	0.01440448
O.I64	Cévní příhoda mozková neurčená jako krvácení nebo infarkt	0.01358037
I27	Jiné kardiopulmonální nemoci	0.01349612
K29	Zánět žaludku a dvanáctníku - gastritis et duodenitis	0.01338366
K62	Jiné nemoci řiti a konečníku	0.01290232
L95	Vaskulitis omezená na kůži, nezařazená jinde	0.01290232
K57	Divertikulární nemoc střeva	0.01217010
I50	Selhání srdce	0.01158408
O.I214	Akutní subendokardiální infarkt myokardu	0.01140944
K80	Žlučové kameny - cholelithiasis	0.01054740

Tab. 1 Veličiny s informačním ziskem větším než 0,01.

Diagnózy s prefixem **O** se u pacienta poprvé objevují až při zkoumané hospitalizaci, ostatní diagnózy jsou převzaty s předchozích hospitalizací pacienta ve vybrané nemocnici během jednoho roku předcházejícímu zkoumané hospitalizaci.

Tyto veličiny jsme použili při učení parametrů modelu logistické regrese. Hodnoty všech veličin jsme normalizovali tak, aby se nacházely v intervalu $<0,1>$. U některých pacientů nebyly všechny vybrané veličiny změřeny. Jednou z možností bylo takové pacienty vynechat. Tím by se ale datový soubor výrazně zredukoval. Proto jsme raději zvolili jednu z metod práce s neúplnými daty, tzv. Multivariate Imputations by Chained Equations [2,3]. Alternativně jsme také otestovali nahrazení chybějící hodnoty rizikového faktoru průměrnou hodnotou tohoto faktoru, ale dosažené výsledky byly horší. Pro vlastní učení parametrů modelu logistické regrese jsme použili funkci *glm*, která je součástí prostředí pro statistické výpočty R [4]. V průběhu výpočtů jsme vyřadili veličiny, které způsobovaly singularity při učení modelu: O.I73, O.I42 a L95 a také ty, které mohly být komplikacemi vzniklými až po přijetí: O.I20, O.I48, O.I46 a O.I46.

Výsledný model je popsán v Tab. 2. V prvním sloupci jsou názvy rizikových faktorů popsaných v Tab. 1. V druhém sloupci jsou jednotlivé koeficienty β – tj. složky vektoru $\boldsymbol{\beta}$ pro vzorec logistické regrese. V třetím sloupci je směrodatná odchylka daného koeficientu. Čtvrtý sloupec obsahuje odpovídající hodnotu $\mathbf{t}$ Studentova t-testu, zda koeficient β má danou střední hodnotu. V pátém sloupci je počet stupňů volnosti Studentova t-rozdělení spočtený podle článku [5]. V posledním šestém sloupci je pravděpodobnost alternativní hypotézy t-testu. Hodnoty nižší než 0.05, což odpovídá hladině statické významnosti 5%, jsou zobrazeny tučně a řádky podbarveny šedou barvou. Tyto hodnoty znamenají, že hypotéza, že koeficient β má udanou hodnotu jako svoji střední hodnotu je přijata na hladině statické významnosti 5%.

	koeficienty β	směrodatná odchylka	hodnota t	stupně volnosti	pravděpodobnost alternativní hyp.
(Intercept)	-2.1060975	1.2651842	-1.664656859	31.598080	0.105868127
S.Urea	8.6381220	3.2418691	2.664549922	6.539030	**0.034363798**
S.kreatinin	-1.0298399	3.1864171	-0.323196842	6.948143	0.756055460
B.leukocyty	1.3401639	2.5917890	0.517080649	6.072126	0.623388122
I48	1.1774437	0.5043932	2.334376623	62.932775	**0.022781215**
O.E78	-1.1969985	0.4437995	-2.697160631	456.217585	**0.007252314**
I73	24.2072289	3127.7478688	0.007739508	463.999987	0.993828154
O.I27	22.4904111	3055.1132840	0.007361564	463.999942	0.994129539
vek	-0.8665293	1.2988153	-0.667169004	46.748401	0.507944173
K92	0.1754608	2.1248919	0.082573969	260.797168	0.934253641
O.I210	2.0831447	1.2175409	1.710944339	14.713859	0.108083943
I74	0.8435731	1.2306952	0.685444349	363.623078	0.493500307
I42	18.6880782	3580.0597075	0.005220047	463.999856	0.995837268
O.I10	-1.7197621	0.5011027	-3.431955378	32.644378	**0.001645515**
O.I211	-18.4366694	1142.1321492	-0.016142326	463.999889	0.987127785
I27	-0.3723272	2.1099515	-0.176462425	220.553439	0.860092594
K29	-0.6493484	1.1797607	-0.550406839	49.675659	0.584507085
K62	1.6111716	3.1727303	0.507818643	83.887081	0.612913195
K57	1.5036105	1.8285737	0.822285931	17.330469	0.422084093
I50	-0.2095659	0.5698730	-0.367741513	18.221241	0.717302894
O.I214	-0.2596627	1.1148068	-0.232921682	9.406510	0.820811800
K80	1.4525334	0.5681741	2.556493236	388.428873	**0.010952816**

Tab 2. Model logistické regrese.

Hodnoty koeficientů **β** se tedy dají zhruba interpretovat následovně. Čím je kladné číslo větší, tím větší je vliv odpovídajícího rizikového faktoru na pravděpodobnost úmrtí. Čím je záporné číslo nižší, tím větší je vliv odpovídajícího rizikového faktoru na pravděpodobnost přežití. Číslo v posledním sloupci nám říká, jak je tento vliv statisticky významný. Pro hodnoty vyšší než $0{,}05$ (což je většina rizikových faktorů) se dá říci, že nebyl vliv na pravděpodobnost úmrtí na našem datovém souboru statisticky prokázán. To je hlavně důsledkem relativně malého počtu pacientů v našem datovém souboru.

6. Výsledky experimentů

Pro spolehlivé vyhodnocení kvality predikce naučeného modelu je třeba nezávislých dat, která nebyla použita k naučení modelu. K tomuto účelu jsme použili metodu *K-násobné křížové validace (angl. K-fold cross-validation)*, kde *K* mělo hodnotu deset. Náš datový soubor jsme náhodně rozdělili do deseti skupin přibližně stejné velikosti. Pro každou z těchto deseti skupin jsme zopakovali následující postup. Zbývajících devět skupin bylo použito pro naučení modelu, který byl na vybrané skupině otestován. Níže uvedené konečné výsledky jsou vždy sumarizací všech deseti dílčích výsledků. Základem pro hodnocení je tzv. *confusion matrix*, která obsahuje:

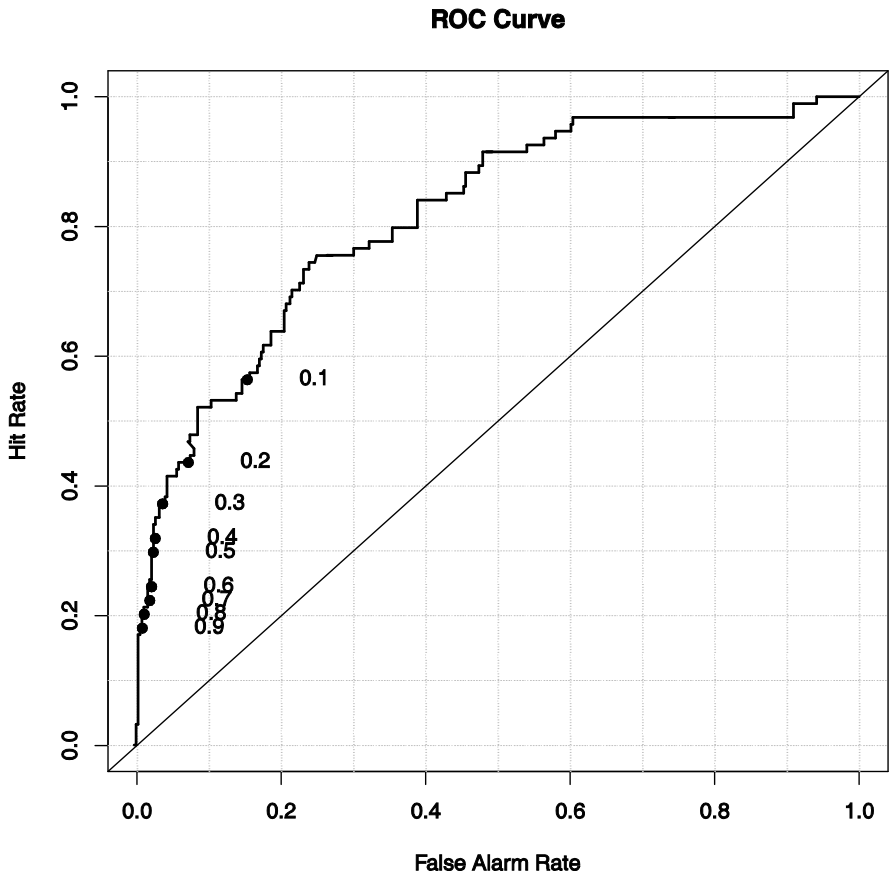
- počet pacientů **tp** („true positives“), kteří byli správně klasifikováni, že zemřou,
- počet pacientů **tn** („true negatives“), kteří byli správně klasifikováni, že nezemřou,
- počet pacientů **fp** („false positives“), kteří byli nesprávně klasifikováni, že zemřou,
- počet pacientů **fn** („false negatives“), kteří byli nesprávně klasifikováni, že nezemřou.

Pro model uvedený v Tab. 2 byly tyto hodnoty **$\mathrm{tp}=28, \mathrm{tn}=383, \mathrm{fp}=9, \mathrm{fn}=66$**. Z toho vychází, že:

- ***úspěšnost*** predikce (angl. *accuracy*) definovaná jako $(\mathrm{tp}+\mathrm{tn})/(\mathrm{tp}+\mathrm{tn}+\mathrm{fp}+\mathrm{fn})$ byla **0,85**,
- ***přesnost*** (angl. *precision, specificity)* definovaná jako $\mathrm{tp}/(\mathrm{tp}+\mathrm{fp})$ byla **0,76**,
- ***úplnost*** (angl. *recall,* nebo *hit rate*) definovaná jako $\mathrm{tp}+/(\mathrm{tp}+\mathrm{fn})$ byla **0,30**,
- ***specificita*** (angl. *specificity* nebo *true negative rate*) definovaná jako $\mathrm{tn}/(\mathrm{tn}+\mathrm{fp})$ byla **0,98** a
- ***falešná pozitivita*** (angl. *false alarm rate)* definovaná jako $\mathrm{fp}/(\mathrm{tn}+\mathrm{fp})$ byla **0,02**.

Výstupem modelu logistické regrese není pouze odhad, zda pacient do 30 dní zemře či nikoliv, ale zároveň model poskytuje pravděpodobnost s jakou k tomu či onomu dojde. V souvislosti s tím je možné měnit práh (který je standardně nastaven na hodnotu $0{,}5$) pro citlivost rozhodnutí, do které skupiny bude pacient klasifikován. Takto je možné například zlepšovat úplnost na úkor přesnosti a naopak. Celkové chování takového klasifikátoru nejlépe charakterizuje tzv. ROC křivka (angl. ROC Curve), viz Obr. 1.

ROC křivka zobrazuje závislost hodnot úplnosti (hit rate) a falešné negativity (false alarm rate) na hodnotě prahu (v obrázku jsou hodnoty prahu uvedeny u bodů křivky). Čím je křivka umístěna výše, tím lepší výsledky model poskytuje. Dobrým měřítkem je velikost oblasti pod křivkou. Tato hodnota je v anglické literatuře obvykle nazývána ROC Area. Maximální hodnotou reprezentující ideální klasifikátor je 1,0. Naopak hodnoty 0,5 dosáhne i náhodný klasifikátor. Hodnota ROC Area našeho modelu byla 0,802.



Obr 1. ROC křivka

7. Závěr

V této práci jsme se zabývali standardizací výsledkového ukazatele „Nemocniční mortalita při akutním infarktu myokardu“. Přestože jsme měli k dispozici relativně malý datový vzorek a ze záznamů pacienta jsme využívali pouze zaznamenané diagnózy a výsledky tří laboratorních testů, podařilo se na základě modelu naučeného na trénovacích datech relativně úspěšně predikovat pro pacienty v testovacích datech, zdali do 30 dní zemřou či nikoliv. Dosažená úspěšnost predikce byla 85% a velikost oblasti pod ROC křivkou byla 0,802. S ohledem na statistické vlastnosti predikčních modelů tohoto typu lze očekávat, že při použití dalších údajů z elektronického záznamu pacienta, jako jsou informace o EKG, lokalizace bolesti, krevní tlak (tyto údaje byly v našem případě zaznamenány pouze ve volném textu a tudíž pro automatické učení byly těžko využitelné), povede k ještě lepší predikci. Pro praktické využití tohoto výsledkového ukazatele je třeba, aby model byl naučen ze záznamů z co největšího počtu nemocnic. Pak také bude možné provést lékařsky odbornou interpretaci výstupů.

Dalším úkolem, kterým bychom se v budoucnu chtěli zabývat, je výzkum společného vlivu kombinací několika rizikových faktorů. Pro to ale bude nezbytné získat větší datový soubor, než jaký jsme měli k dispozici.

Poděkování

Tato práce je výsledkem spolupráce Fakulty managementu VŠE Praha a firmy STAPRO, s.r.o. v rámci projektu Zimolez číslo 2C06019 poskytnutého MŠMT ČR.

Kontaktní adresy

Václav Kratochvíl, Jiří Vomlel, Petr Somol
Ústav teorie informace a automatizace AV ČR, v.v.i
Pod vodárenskou věží 4
182 08 Praha 8 – Libeň
a
Fakulta managementu
Vysoká škola ekonomická v Praze
Jarošovská 1117/II
377 01 Jindřichův Hradec

Hynek Kružík, Petr Tůma
Gnomon, s.r.o.
Faltysova 1500/18
15600 Praha 5 - Zbraslav

Literatura

[1] Harlan M. Krumholz, Sharon-Lise T. Normand, Deron H. Galusha, Jennifer A. Mattera, Amy S. Rich, Yongfei Wang, Yun Wang. *Risk-Adjustment Models for AMI and HF 30-Day Mortality, Methodology*. Harvard Medical School, Department of Health Care Policy, 2007.

[2] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, 1987.

[3] S. Van Buuren, J.P.L Brand, C.G.M. Groothuis-Oudshoorn, D.B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 12, 1049–1064, 2006.

[4] R Development Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0, http://www.R-project.org

[5] J. Barnard and D.B. Rubin. Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955, 1999.