PREDIKCE PRAVDĚPODOBNOSTI PŘEŽITÍ CESTUJÍCÍCH NA TITANICU

Teoretické ukotvení

Potopení zaoceánského parníku *R.M.S. Titanic* patří mezi celosvětově známé tragédie. Ve své době šlo o největší parník světa. Byl určen pro převoz cestujících mezi Evropou a Severní Amerikou. Kapacita lodi mohla převážet až 2 603 cestujících, včetně kočárů a automobilů. O provoz lodě a o pohodlí cestujících se staralo až 899 členů posádky (wikipedie, nedat.). 15. dubna 1912 se během své první plavby Titanic potopil po srážce s ledovcem. Tehdy zemřelo 1502 z 2224 cestujících včetně posádky. Tato tragédie šokovala mezinárodní společenství a vedla k lepším bezpečnostním předpisům pro lodě (Seghal, 2018).

Jedním z důvodů, proč zde byla tak vysoká ztráta na životech, byl nedostatek záchranných člunů pro cestující a posádku. Některé skupiny lidí však měly větší šanci přežít než jiné. Čtyři roky stará studie *„Exploring the Titanic Dataset“* provedla analýzu toho, kteří cestující měli větší šanci na přežití (Srini, 2019).

Statistická metoda

Na odhad pravděpodobnosti přežití cestujících jsme použili **binární logistický regresní model**. Jedná se o statistický model, který popisuje chování dichotomické závislé proměnné. Výstup nabývá pouze dvou hodnot za pomocí spojitých či kategoriálních vysvětlujících proměnných. V naší zprávě si klademe za cíl prozkoumat, který z prediktorů má největší vliv na přežití pasažérů při potopení parníku Titanic.

Datový soubor

Srini (2019) ve své explorativní analýze použil rozsáhlý datový soubor, který čítal 1305 respondentů. Náš původní datový soubor měl 1314 respondentů, nicméně jsme museli data očistit z důvodu chybějících údajů. Konečný datový soubor má 756 mužů a žen různého věku.[1] V datové matici je zastoupeno 288 (38,1 %) žen a 468 (61,9 %) mužů. Respondenti byli dále rozděleni do tří tříd na palubě. V první třídě bylo 226 (29,9 %) cestujících, ve druhé 212 (28 %) a ve třetí 318 (42,1 %). V tabulce č. 1 jsou uvedeny základní popisné charakteristiky.[2]

[1] Data a další informace o této zprávě jsou dostupné na https://dostal.vyzkum-psychologie.cz/stat4/zprava.php?id=29.

[2] Data k této zprávě byly poskytnuty kolegy z katedry psychologie na Masarykově univerzitě v Brně.

Tabulka 1: Popisné charakteristiky

Prediktor	Průměr	Minimum	Maximum	Směrodatná odchylka
Věk	30,4	0,17	71	14,26
Pohlaví	101,62	101,62	102	0,49
Třída	102,12	101	103	0,84

Závislé proměnné a regresory

Naší závislou proměnnou je **přežití**. Kladná odpověď je kódovaná jedničkou, záporná odpověď je kódována nulou. Zvolené prediktory pravděpodobnosti jsou následující:

- **Věk** – nejstarší cestující – 71 let, nejmladší cestující 2 měsíce, průměrný věk cestujících – 30,4 let
- **Pohlaví** – viz. výše
- **Třída cestujících – 1., 2. a 3. třída**, počet cestujících – viz. výše

Pro výpočet logistické regrese byl použit program *Statistica*. Pro ověření statistické významnosti regresního koeficientu jsme použili Waldovu statistiku, která má v případě logistické regrese za planosti nulové hypotézy normované normální rozdělení. Výsledky testů shrnuje tabulka 2, která ukazuje odhadnuté efekty několika prediktorů (věk, pohlaví a třída) spolu s hladinou významnosti p, poměru šancí a regresními koeficienty.[3]

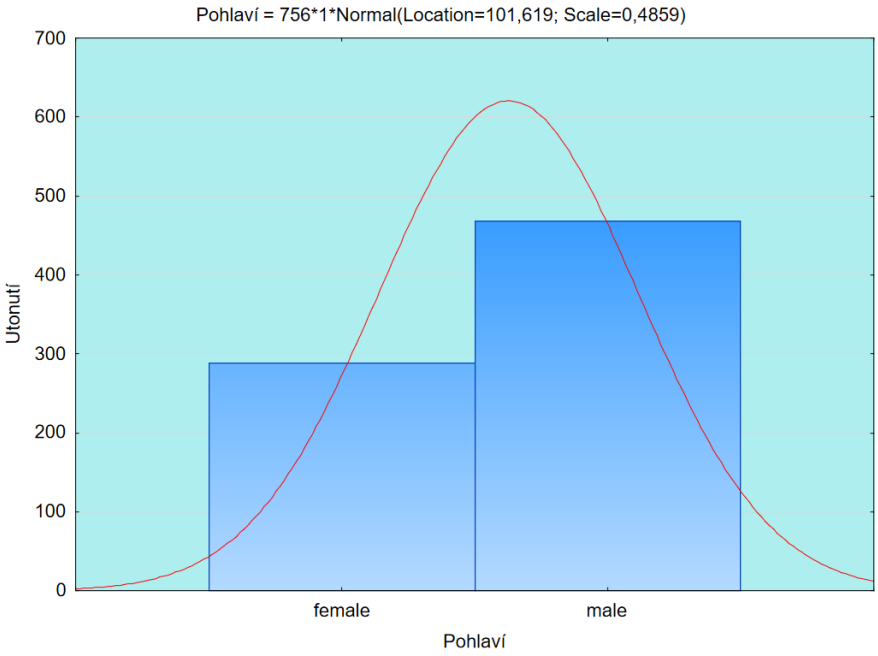
Tabulka 2: Výsledky testů statistické významnosti

Prediktory	Regresní koeficient	Testová statistika	Šance	p hodnota
Konstanta	-1,4	29,8		
Věk	-0,04	26,46	0,96	<0,00
Pohlaví	2,63	170,52	13,9	<0,00
1. Třída	2,52	83,03	12,45	<0,00
2. Třída	1,23	26,83	3,42	<0,00

Z tabulky je patrné, že nulová p hodnota značí u všech regresorů statistickou významnost. Všechny tedy mají vliv na pravděpodobnost přežití. Jako nejvýznamnější prediktor se ukázalo být **pohlaví**, což lze vyčíst z výsledků testové statistiky i poměru pravděpodobností (šance). Tedy platí, že ženy mají téměř o 14x vyšší šanci na přežití než muži. To je graficky znázorněno v histogramu č. 1.

[3] Konstanta značí referenční skupinu, a to muže, kteří cestovali 3. třídou

Graf 1: Prediktor přežití v závislosti na pohlaví



Pro **třídu cestujících** platí, že oproti 3. třídě mají pasažéři cestující v 1. třídě přibližně 12x větší šanci na přežití a ve druhé třídě mají šanci 3,4x vyšší. Poslední závislou proměnnou je **věk**, kdy je patrné, že s každým přibývajícím rokem se šance na přežití sníží o necelé 4 %.

Kvalitu predikce našeho modelu jsme ověřili prostřednictvím ukazatele Nagelkerke R 2. Model se všemi regresory vysvětluje 47,69 % rozptylu (Nagelkerke R2 = 0,4769), viz. tabulka č. 3.

Tabulka 3: Ukazatelé kvality modelu

	Hodnota	**Hodnota v %**
Cox-Snell R2	0,35	35 %
Nagelkerke R2	0,48	48 %

Závěr

Z výsledků parametrů odhadu lze vidět statistickou významnost mezi všemi třemi prediktory, stejně jako je tomu ve studii Srini (2019). Věk je negativně korelován s pravděpodobností přežití. Co se týče pohlaví, tak se ukázalo, že ženy měly až 14x větší šanci na přežití oproti mužům. Pro pasažéry v první třídě se ukázala pravděpodobnost přežití jako nejvyšší z celého datového souboru.

Seznam zdrojů

- Kaggle (2012). *„Titanic: Machine Learning from Disaster.“* Dostupné z: https://www.kaggle.com/competitions/titanic/overview
- Seghal, M. (2018). *„Titanic Data Science Solution.“* Kaggle. Dostupné z: https://www.kaggle.com/code/startupsci/titanic-data-science-solutions/notebook
- Srini, A. (2019). *„Exploring the Titanic Dataset.“* Towards Data Science. Dostupné z: https://www.kaggle.com/code/shashaalam/exploring-the-titanic-dataset
- Wikipedie (nedat). *Titanic*. Dostupné z: https://cs.wikipedia.org/wiki/Titanic