Posudek školitele diplomové práce

Studentka: **Bc. Barbora Pánková**

Název práce: **Analýza cen pojištění pomocí strojového učení**

Autorka v práci zkoumá a srovnává možnosti využití několika různých přístupů pro odhadování cen konkurentů v pojištění odpovědnosti z provozu vozidla na základě omezených sad pozorování těchto cen, přičemž přístupy jsou založené na využití zobecněných lineárních modelů a modelů neuronových sítí. To je na jednu stranu praktický problém, který pojišťovny různými metodami řeší, na druhou stranu je to problém obtížný z důvodu neveřejné cenotvorby a omezené dostupnosti dat na jedné straně a na druhé straně z důvodu komplexní struktury cen, kdy pojišťovny nabízejí rozdílné ceny nejen napříč segmenty klientů, ale i napříč prodejními kanály, na základě schopnosti klienta vyjednávat o ceně, sezonních slev a dalších proměnných faktorů. Z toho důvodu nelze doufat v jedno univerzální, „teoreticky vždy správné" řešení problému a praktické zkoušení a porovnávání různých přístupů tedy považuji za vhodný přístup.

Autorka dle mého názoru splnila všechny body zadání: seznámila se s modelováním pomocí metod GLM a dále i neuronových sítí, sestavila pomocí těchto metod několik alternativních přístupů pro odhad cen na větším vzorku vstupních dat (v práci označován jako pojistitel A/bílý pojistitel) i na jiném, velmi omezeném vzorku dat (pojistitel B/černý) a porovnala, které přístupy spolehlivěji predikují pozorované ceny daného pojistitele.

V práci bych vyzdvihl zejména praktickou stránku aplikace použitých metod na konkrétní problém. Autorka se dobře vyrovnala s tím, že v teoretické rovině jsou zvolené metody v literatuře sice dobře popsané, ale zdroje popisující jejich aplikaci na praktické problémy tohoto typu jsou dostupné jen v omezené míře, a dokázala samostatně nalézt přístupy vhodné pro specifickou podobu vstupních dat. To zahrnuje přístupy k počáteční transformaci mnohorozměrných a heterogenních dat (mix kategorických vstupů, spojitých vstupů typu věk, vícerozměrných spojitých vstupů typu geografická lokace či mnohorozměrných kategorických vstupů jako např. typ a veškeré vlastnosti vozu), předkategorizaci a předvýběru proměnných a jejich kategorií (zejm. v případě GLM) nebo k výběru konkrétního modelu z řady alternativ (zahrnutí/nezahrnutí určitých proměnných a kategorií do modelu, výběr typu GLM modelu, výběr struktury neuronové sítě a nastavení kalibračních epoch atd.). Zejména v případě využití neuronových sítí právě potřebu zvládnutí těchto „přípravných prací" považuji za jeden z důvodů, proč jsou tyto metody v praxi českých pojišťoven dosud daleko méně využívané než např. GLM, a práci tedy v tomto bodě považuji za inovativní. Věřím, že zvládnutí těchto přístupů autorku dobře kvalifikuje k řešení i řady jiných praktických úloh prediktivního modelování.

Závěry takto prakticky pojaté práce jsou omezeny konkrétní strukturou použitých datových sad. Jak autorka v závěru sama poznamenává, oproti datovým sadám dostupným pro vytvoření práce jsou dnes již ceny na trhu strukturované hlouběji, zejména z hlediska škodní historie klienta. Přesto osobně očekávám, že následující autorčiny závěry se ukáží jako obecně platné: u velmi omezeného vzorku příliš nepomohou ani pokročilejší metody, u dostatečného vzorku naopak dává dobré výsledky vícero přístupů, přičemž mezi silné stránky GLM patří snadná interpretace výsledných faktorů, zatímco u neuronových sítí se při dobrém zvládnutí výše uvedených „přípravných prací" může jako jejich silná stránka ukázat větší volnost a rychlost prvotní analýzy.

Na základě výše uvedeného jsem přesvědčen, že autorka zadání své diplomové práce splnila, práci doporučuji k obhajobě a navrhuji ji hodnotit známkou A (výbornou).

V Praze, dne 7. července 2020

Tomáš Petr
školitel