



Hodnocení vedoucího závěrečné práce

Vedoucí práce: Ing. Jakub Marian, M.Sc.
Student: Bc. Peter Kolárovec
Název práce: Rekonstrukce pseudosociálních sítí z bankovních transakčních dat
Obor / specializace: Znalostní inženýrství
Vytvořeno dne: 14. ledna 2023

Hodnotící kritéria

1. Splnění zadání

- ▶ [1] zadání splněno
- [2] zadání splněno s menšími výhradami
- [3] zadání splněno s většími výhradami
- [4] zadání nesplněno

Autor poskytuje k zadání obecný kontext a následně se zabývá implementací reálného use case v rámci České spořitelny, čímž zcela splnil zadání.

2. Písemná část práce

80/100 (B)

V první části práce autor stručně popisuje principy zpracování bankovních transakčních dat, které jsou ve velkých bankách již standardem. Ve druhé části stručně popisuje příklady různých vztahů (hran) v rámci sociálních sítí. Po faktické stránce nelze těmto částem nic vytknout, ocenil bych větší míru provázanosti s následujícími částmi (motivace podávaných informací je bez přečtení zbytku práce poněkud nejasná).

Ve třetí části autor představuje framework SEAL a správně usuzuje, že v případě bankovních klientů je vhodné ho obohatit o hranové features založené na bankovních transakcích.

Ve čtvrté části autor popisuje proces CRISP-DM pro vývoj modelu, jehož rozšířená varianta je používána i interně v České spořitelně, a aplikuje ho na byznysový problém identifikace domácností. V rámci byznysové motivace spočítal, že 68 % domácností identifikovaných současným heuristickým algoritmem jsou jednočlenné domácnosti, což považuje za podezřele vysoké číslo. Zde by však bylo nejprve vhodné provést analýzu, kolik jednočlenných domácností bychom měli mezi klienty ČS očekávat v rámci jejího market share s použitím statistických dat o velikostech domácností v ČR a známé demografie klientů ČS.

Autor dále popisuje preprocessing dat, rozdělení na trénovací, validační a testovací množinu a řeší problémy specifické pro daný dataset, totiž že předem známý předpoklad úplnosti grafů může vést k target leakage. Popisuje tvorbu jednotlivých features, bylo by ovšem vhodné věnovat více prostoru popisu implementace algoritmu node2vec.

V částech 4.4 a 4.5 a v části 5 se autor věnuje trénování neuronové sítě (DGCNN), vyhodnocení modelu, jehož cílem je připojit jednočlenné domácnosti k vícečlenným, a následné rekonstrukci celé "domácnostní" sítě. Zde by bylo vhodné srovnat výsledky s jednoduchým modelem jakožto benchmarkem výkonnosti, např. s pomocí agregace features přes celou potenciální domácnost a využití v ČR běžně používaného ML modelu (např. XGBoost), jelikož v případě domácností je grafová struktura velice jednoduchá (všechny komponenty jsou úplné grafy) a lze očekávat, že přínos modelu využívajícího grafovou topologii nebude oproti jednodušším třídám modelů velký. Otevřeným problémem také zůstává spojení dvou jednočlenných domácností do jedné dvoučlenné.

3. Nepísemná část, přílohy

85 /100 (B)

Vypracované kódy v jazyce Python (Jupyter notebooky) splňují očekávané standardy v rámci analytické práce v oddělení Data Science České spořitelny a svým rozsahem i obsahem výrazně překračují běžně realizované projekty, takže z tohoto hlediska hodnotím na výbornou. Nejsou však nijak členěné (ačkoliv to formát Jupyter přirozeně umožňuje) a neobsahují žádné komentáře i přesto, že místy obsahují na první pohled hůře srozumitelné části s hardcodovanými indexy či složitější větvící logikou, což ubírá kódu na čitelnosti. Také bych docenil komentáře upozorňující na problematická místa, která by při budoucím vývoji vyžadovala zvýšenou pozornost.

4. Hodnocení výsledků, jejich využitelnost

95 /100 (A)

Práci považuji za proof of concept využití grafových neuronových sítí v rámci České spořitelny, jedná se o první projekt tohoto typu v rámci celé společnosti. Ačkoli toto konkrétní business zadání by bylo možné řešit i jinými metodami, popsané a implementované grafové metody mají velký potenciál pro širokou řadu analytických problémů, kterými se Česká spořitelna může v budoucnosti zabývat, a přínos získaného know-how je nesporný.

5. Aktivita studenta

- [1] výborná aktivita
- [2] velmi dobrá aktivita
- ▶ [3] průměrná aktivita
- [4] slabší, ale ještě dostatečná aktivita
- [5] nedostatečná aktivita

6. Samostatnost studenta

- ▶ [1] výborná samostatnost
- [2] velmi dobrá samostatnost
- [3] průměrná samostatnost
- [4] slabší, ale ještě dostatečná samostatnost
- [5] nedostatečná samostatnost

Celkové hodnocení

85 /100 (B)

Práce splnila očekávání a má zřejmý praktický přínos pro autora zaměstnavatele. Hlavní výtka je přílišná stručnost. Mnoho myšlenek by bylo vhodné dále rozvést a ukázat podpůrné analýzy vedoucí ke konkrétním rozhodnutím v rámci implementace, ale celkově práci hodnotím jako zdařilou.

Instrukce

Splnění zadání

Posudte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posudte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posudte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Aktivita studenta

V souvislosti s průběhem a výsledkem práce posudte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven.

Samostatnost studenta

V souvislosti s průběhem a výsledkem práce posudte schopnost studenta samostatně tvůrčí práce.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.