ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta Elektrotechnická

Basics of Data Analysis

Semestral Work

SIT - BI

Volha Kramko

2025

Situation

V rámci semestrální práce pro předmět Základy datové analýzy mě zaujalo téma popularity knih. Rozhodla jsem se analyzovat rozsáhlý dataset z platformy Goodreads, který obsahuje metadata knih a hodnocení od čtenářů. Cílem bylo zjistit, jaké faktory ovlivňují počet pozitivních recenzí a obecnou oblíbenost knih.

Task

Cílem bylo pomocí datové analýzy a regresního modelu prozkoumat vztahy mezi různými charakteristikami knih — jako je délka, jazyk, rok vydání, autor či nakladatel — a jejich popularitou. Konkrétním úkolem bylo zjistit, které faktory mají největší vliv na čtenářské hodnocení a vytvořit model, který by popularitu predikoval.

Action

Data byla získána z veřejného datasetu a rozdělena do dvou hlavních částí: metadata knih (název, autor, rok, počet stran) a uživatelská hodnocení (textová i kvantitativní). Nejprve proběhlo čištění a sloučení dat, následně agregace a kategorizace proměnných. Následovala vizualizace trendů a tvorba lineárního regresního modelu, kde cílovou proměnnou byl počet pozitivních recenzí (logaritmicky transformovaný).

Result

Z výsledků analýzy vyplynulo, že na popularitu knih mají pozitivní vliv zejména tyto faktory: známý autor, anglický jazyk, delší rozsah knihy a moderní rok vydání. Regresní model vysvětlil přibližně 22 % variability, což je akceptovatelné vzhledem k povaze dat. Největší pozitivní vliv měly knihy psané anglicky a vydané renomovanými nakladatelstvími.

Review

Analýza má smysl a odpovídá očekáváním. Lze ji dále rozšířit o práci s textovými popisy knih (např. NLP analýza anotací) a zahrnout více jazykových a kulturních kontextů. Limitací je, že průměrné hodnocení a počet recenzí ovlivňují i externí faktory, jako je popularita autora v médiích či filmové adaptace. Do budoucna by bylo vhodné provést hlubší analýzu jednotlivých žánrů a využít pokročilejší modely jako např. Random Forest nebo XGBoost.