

Technicka sprava: Regresny model pre predikciu

1. Teoreticky opis pouziteho regresneho modelu

V tomto projekte sme pouzili linearnu regresiu ako nas regresny model. Linearna regresia je statisticka metoda, ktora sa pouziva na modelovanie linearného vzťahu medzi zavislou premennou a jednou alebo viacerými nezávislými premennými. Model predpoklada, že vzťah medzi premennými moze byt vyjadreny linearnou funkciou.

Matematicky mozeme linearnu regresiu vyjadrit ako: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$

kde y je zavisla premenna, x_1, x_2, \dots, x_n su nezavisle premenne, β_0 je intercept (hodnota y , ked vsetky x su 0), $\beta_1, \beta_2, \dots, \beta_n$ su koeficienty (vahy) pre kazdu nezavislu premennu, a ε je chybovy clen.

2. Metodologia riesenia

Pre implementáciu lineárnej regresie sme zvolili knižnicu scikit-learn v Pythone. Tento vyber bol motivovaný niekoľkými faktormi:

a) Jednoduchosť použitia: scikit-learn poskytuje intuitívne API pre trenovanie a vyhodnocovanie modelov. b) Výkonnosť: Implementácia v scikit-learn je optimalizovaná a efektívna. c) Integrácia: Lahko sa integruje s ďalšími nástrojmi pre manipuláciu s dátami, ako numpy a pandas.

Proces implementácie zahrňal nasledujúce kroky: 1. Nacítanie dát pomocou numpy. 2. Predspracovanie dát - konverzia kategorických premenných pomocou LabelEncoder. 3. Rozdelenie dát na trenovaciu a evaluacnú množinu. 4. Trenovanie modelu lineárnej regresie na trenovacích dátach. 5. Vyhodnotenie modelu pomocou R^2 skóre. 6. Predikcia na evaluacných dátach a uloženie výsledkov.

3. Diskusia o výsledkoch

Nas model dosiahol R^2 skóre 0.8236 na trenovacích dátach, čo naznačuje, že model vysvetľuje približne 82.36% variability v cieľovej premennej. Toto je pomerne dobrý výsledok, ktorý naznačuje, že model zachytil významné vzory v dátach.

Avšak je dôležité poznamenať, že toto skóre bolo dosiahnuté na trenovacích dátach, čo moze viesť k nadhodnoteniu skutočnej výkonnosti modelu. Pre robustnejšie hodnotenie by bolo vhodné použiť krízovú validáciu alebo samostatnú validacnú množinu.

Ďalšie prístupy, ktoré by mohli byť vyskúšané pre potenciálne zlepšenie výkonu modelu, zahrňajú: 1. Použitie pokročilejších regresných techník, ako je Ridge alebo Lasso regresia, ktoré mozu pomôcť s regularizáciou. 2. Experimentovanie s rôznymi metodami predspracovania dát, napríklad skalovaním funkcií. 3. Skúmanie nelineárnych vzťahov v dátach a potenciálne použitie polynomiálnych funkcií.

Záver: Nas lineárny regresný model dosiahol slubné výsledky na trenovacích dátach. Avšak pre komplexnejšie hodnotenie a potenciálne zlepšenie výkonu by bolo vhodné implementovať ďalšie techniky a vykonať dôkladnejšiu analýzu na validacnej množine dát.