Teoretická časť

Regresný model použitý v riešení je **Lasso regresia** (Least Absolute Shrinkage and Selection Operator). Ide o lineárny model, ktorý minimalizuje sumu štvorcov odchýlok medzi predikovanými a skutočnými hodnotami cieľovej premennej, pričom pridáva penalizačný člen na zjednodušenie modelu. Penalizačný člen závisí od absolútnych hodnôt váh modelu, čo má za následok vynulovanie niektorých koeficientov, čím dochádza k výberu relevantných vlastností (feature selection).

Výhodou Lasso regresie je jej schopnosť znížiť riziko pretrénovania a zvýšiť interpretovateľnosť modelu. Lasso je obzvlášť vhodné pri vysokorozmerných dátach alebo v prípadoch, keď sa očakáva, že niektoré vlastnosti budú irelevantné.

Návrh riešenia

Riešenie pozostávalo z niekoľkých krokov, ktoré zahŕňali predspracovanie dát, optimalizáciu modelu a hodnotenie jeho výkonnosti:

1. **Predspracovanie dát**:
   - Dátová množina obsahovala **kategorické** a **numerické** stĺpce. Prvých 10 stĺpcov bolo považovaných za kategorické.
   - Numerické dáta boli spracované pomocou **imputácie mediánom** a následne škálované pomocou **MinMaxScaler**.
   - Kategorické dáta boli spracované pomocou **imputácie najčastejšou hodnotou** a zakódované pomocou **OneHotEncoder**.
2. **Optimalizácia hyperparametrov**:
   - Na optimalizáciu hyperparametrov modelu bola použitá metóda **GridSearchCV** s krosvalidáciou (5-fold cross-validation). Pre Lasso regresiu sa testovali rôzne hodnoty regularizačného parametra alpha a maximálny počet iterácií.
3. **Výber najlepšieho modelu**:
   - Na hodnotenie výkonnosti modelu bol použitý **R^2 skóre**. Model bol testovaný na rôznych náhodných rozdeleniach dát pomocou train_test_split s rôznymi hodnotami random_state.
   - Výsledky jednotlivých modelov boli zoradené podľa dosiahnutého R^2 skóre a bol vybraný najlepší model na základe týchto metrík.
4. **Predikcia na evaluačných dátach**:
   - Po natrénovaní najlepšieho modelu boli predikcie na evaluačných dátach uložené do súboru **y_predikcia.npy** vo formáte numpy, aby vyhovovali zadaniu.

Diskusia a výsledky

Počas riešenia bolo testovaných niekoľko prístupov a rôzne hodnoty náhodného stavu (random_state). Výsledky modelov s najvyššími R^2 skóre sú nasledovné:

Model	Random State	R^2 skóre
Lasso	**123**	**0.983781**
Ridge	**123**	**0.983299**
LinearRegression	**123**	**0.983067**
ElasticNet	**123**	**0.982030**
Lasso	**42**	**0.976700**

Model **Lasso s random_state=123** dosiahol najvyššie R^2 skóre 0.983781. Tento model bol použitý na predikciu evaluačných dát.

Výhody zvoleného prístupu:

- **Lasso regresia** umožňuje výber relevantných vlastností, čo zlepšuje interpretáciu modelu.
- Použitie GridSearchCV umožnilo efektívne nájsť najlepšie hodnoty hyperparametrov.
- Predspracovanie dát bolo navrhnuté tak, aby zvládlo chýbajúce hodnoty a kombináciu rôznych typov údajov.

Limity a poznámky:

- Model je citlivý na výber regularizačného parametra alpha. Nevhodné nastavenie môže viesť k zhoršeniu výkonnosti.
- Práca s evaluačnými dátami neumožňuje presné vyhodnotenie na dátach mimo trénovacej množiny.

Záver

Cieľom bolo navrhnúť a implementovať regresný model na predikciu hodnôt cieľovej premennej s použitím evaluačných dát. Lasso regresia sa ukázala ako vhodná voľba, dosahujúc vysoké R^2 skóre na trénovacích dátach. Model bol úspešne aplikovaný na evaluačné dáta a predikcie boli uložené vo formáte podľa zadania.

V budúcnosti by bolo možné experimentovať s inými modelmi (napr. Random Forest alebo Gradient Boosting) a alternatívnymi stratégiami predspracovania dát, aby sa zistilo, či je možné dosiahnuť ešte vyššiu presnosť.