Procena cene kuća pomoću neuronske mreže

Definicija problema

Cilj projekta je razviti regresioni model koji može, što preciznije proceniti cenu kuća na osnovu karakteristika nekretnine (poput površine/kvadrature, broja soba, godine izgradnje…). Problem se rešava regresijom pri čemu je ciljno obeležje - sama cena kuće.

Motivacija

Procena cena kuća (nekretnina) ima praktičnu primenu u industriji nekretnina i finansijama:

- pomoć agentima i kupcima da odrede što verodostojniju tržišnu vrednost nekretnine
- omogućava bankama i kreditnim institucijama bolje procene rizika pri odobravanju kredita
- doprinosi analizi tržišta i planiranju investicija

Skup podataka

Za potrebe projekta odabran je “Ames Housing Data Set”, poznat i široko korišćen u zadacima regresije unutar oblasti mašinskog učenja.

Link do skupa podataka:
https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset

Skup podataka sadrži **2930 zapisa (kuća)** i **80 atributa**, koji obuhvataju različite karakteristike nekretnina.
Primeri značajnih atributa za procenu cene kuća:

- **OverallQual** - kvalitet materijala i izrade
- **GrLivArea** - površina iznad zemlje (površina stambenog prostora)
- **TotalBsmtSF** - ukupna površina podruma
- **YearBuilt** - godina izgradnje
- **LotArea** - površina placa
- **BedroomAbvGr** - broj spavaćih soba iznad podruma
- **FullBath, HalfBath** - broj kupatila
- **GarageArea** - veličina garaže

Ciljno obeležje je “SalePrice”, odnosno “prodajna cena”, koja je izražena u američkim dolarima (USD) čije su vrednosti numeričke.

Način pretprocesiranja podataka

Pre uvodjenja u model, podaci će biti pripremljeni na sledeći način:

- **Uklanjanje i imputacija nepotpunih zapisa i atributa**: atributi zapisa koji imaju previše nedostajućih vrednosti biće uklonjeni. Kod atributa koji imaju manji broj nedostajućih vrednosti biće primenjena imputacija, tačnije, nedostajuće vrednosti biće zamenjene srednjom vrednošću (ili medijanom ukoliko bude bilo prisustvo outlier-a) za atribute sa numeričkim vrednostima i najčešćom vrednošću za kateogrijske atribute (Neighbourhood, HouseStyle…).
- **Kodiranje kategorijskih promenljivih**: svi tekstualni atributi biće pretvoreni u numeričke vrednosti korišćenjem **label encoding-a**, kako bi mogli biti procesirani.
- **Analiza korelacije atributa sa ciljnim obeležjem**: iz skupa će biti uklonjeni atributi koji pokazuju veoma nisku povezanost sa ciljnim obeležjem (SalePrice), čime se smanjuje prisustvo nerelevantnih atributa i olakšava treniranje modela.
- **Smanjenje dimenzionalnosti primenom PCA algoritma**: nakon izbora značajnih atributa, primeniće se Principal Component Analysis (PCA) radi dodatnog smanjenja dimenzionalnosti i transformacije postojećih u novi skup atributa, uz očuvanje što većeg dela varijanse podataka.

Metodologija

Proces rešavanja problema procene cena nekretnina (kuća) se sastoji iz nekoliko faza:

- **Priprema podataka**: obuhvata učitavanje podataka kao i sve korake njhovog pretprocesiranja poput uklanjanja nepotpunih zapisa i atributa, imputacije nedostajućih vrednosti, redukovanje dimenzionalnosti skupa atributa.
- **Podela podataka**: obuhvata podelu podataka na trening/validacioni/test skup, radi treniranja modela, podešavanja hyperparametar-a i konačne evaluacije.
- **Treniranje modela**: obuhvata izgradnju neuronske mreže za regresiju. Model će imati **ulazni sloj** odgovarajuće dimenzije za broj atributa, nekoliko **skrivenih slojeva** i **izlazni sloj** sa jednim neuronom za procenu cene kuće. Tokom treniranja koristiće se **MSE funkcija gubitka** i **Adam optimizer** za ažuriranje težina.
- **Validacija i optimizacija**: obuhvata praćenje performansi na validacionom skupu tokom podešavanja hiperparametara (npr. broj slojeva neurona, learning rate) i odabir najbolje konfiguracije na osnovu metrika preformansa.
- **Evaluacija i testiranje**: Konačna procena modela nad test skupom, analiza rezultata kroz vizuelizaciju grešaka i poredjenja procenjenih i stvarnih vrednosti.

Način evaluacije

Rezultati modela će se evaluirati pomoću standardnih metrika za regresiju. Podaci će biti podeljeni na **trening, validacioni i test skup** – u odnosu **70:15:15**, kako bi se omogućila optimizacija hiperparametara na validacionom skupu i konačna provera performansi na test skupu.

Za merenje performansi koristiće se sledeće metrike:

- **Mean Squared Error** - srednja kvadratna greška između predviđenih i stvarnih cena kuća
- **Root Mean Squared Error (RMSE)** - koren srednje kvadratne greške, čime je omogućena interpretacija u istim jedinicama kao i ciljna promenljiva (USD)
- **Mean Absolute Error (MAE)** – prosečna apsolutna greška, koja je manje osetljiva na ekstremne vrednosti nego MSE

Tehnologije

Tehnologije koje će biti korišćene u razvoju modela:

- **Python** - glavni programski jezik za implementaciju modela
- **Pandas** - manipulacija i analiza podataka (učitavanje CSV-a, rad sa DataFrame-ovima).
- **NumPy** - numeričke operacije i rad sa matricama.
- **Scikit-learn** - za pretprocesiranje podataka (imputacija nedostajućih vrednosti, enkodiranje kategorijskih atributa…), podelu podataka na train/val/test i evaluaciju modela.
- **PyTorch** - za izgradnju i treniranje neuronske mreže za regresiju.
- **Matplotlib / Seaborn** - za vizualizaciju podataka, distribucije atributa i grešaka modela.

Relevantna literatura

- “House Price Prediction Using Machine Learning for Ames, Iowa“: studija koja istražuje primenu različitih modela, uključujući i neuronsku mrežu, za predikciju cenu kuća u Ames-u (https://www.researchgate.net/publication/382199393_House_price_prediction_using_machine_learning_for_Ames_Iowa)
- “Using Artificial Neural Networks for Regression in Python”: članak o implementaciji duboke neuronske mreže za regresiju u Pythonu, autor koristi skup podataka o starim automobilima, ali metodologija je primenljiva i na procenu cena kuća (https://thinkingneuron.com/using-artificial-neural-networks-for-regression-in-python/)