Анализа стабилности и поузданости модела линеарне регресије на реалним подацима

1. Опис проблема

Линеарна регресија представља једну од основних метода за моделовање односа између зависне и независних променљивих и често се користи у анализи реалних података. Међутим, у пракси се подаци ретко понашају идеално. Често садрже шум, екстремне вредности и међусобно повезане атрибуте, што може довести до нестабилних параметара и непоузданих резултата.

Класична метода најмањих квадрата (OLS) позната је по томе што је осетљива на овакве неправилности у подацима. И мале промене у улазним подацима могу довести до значајних промена у вредностима регресионих коефицијената, што утиче на интерпретабилност и применљивост модела. У оквиру овог пројекта нагласак је стављен на самосталну имплементацију различитих метода линеарне регресије и на детаљно разумевање њиховог рада у позадини, начина оптимизације и међусобних разлика. Моделовање цене станова користи се искључиво као илустративни пример на коме се ове методе примењују и упоређују.

Циљ овог пројекта је да се испита стабилност и поузданост различитих модела линеарне регресије у условима реалних, шумних података. Посебан акценат ставља се на анализу понашања параметара модела и њихове осетљивости на промене у скупу података, а не искључиво на тачност предвиђања. На конкретном примеру цена некретнина промене се често могу јавити услед одступања цене код некретнина које су за адаптацију, неукњижених некретнина, продате испод цене…

---

2. Скуп података

За реализацију пројекта користиће се скуп података House Prices: Advanced Regression Techniques, доступан на платформи Kaggle:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Основне карактеристике скупа података:

- 1460 инстанци
- 79 атрибута

Зависна променљива:

- **SalePrice** – продајна цена куће

Изабране независне променљиве:

- **GrLivArea** – површина стамбеног простора
- **OverallQual** – укупан квалитет објекта
- **YearBuilt** – година изградње

- **TotalBsmtSF** – површина подрума
- **GarageArea** – површина гараже

Ове променљиве су одабране јер имају јасно значење, добру интерпретабилност и значајан утицај на цену некретнина.

---

3. Методологија

3.1 Припрема података

У првој фази биће извршена припрема података која обухвата:

- обраду недостајућих вредности (импутацију),
- кодирање категоријских променљивих,
- нормализацију нумеричких атрибута.

Након припреме, скуп података ће бити подељен на тренинг и тест део у односу 80/20.

3.2 Основни модел

Као референтни модел користиће се вишеструка линеарна регресија са OLS методом. Овај модел ће служити као полазна тачка за поређење са осталим приступима и омогућити анализу утицаја неправилности у подацима на параметре модела.

3.3 Анализа утицаја шумних података

Како би се симулирали реални услови, у податке ће бити контролисано унет мањи ниво шумних вредности. Овај поступак ће омогућити анализу:

- промена у вредностима регресионих коефицијената,
- стабилности модела при малим изменама у подацима,
- утицаја ових промена на резултате предвиђања.

3.4 Избор метода

Методе које ће бити примењене изабране су тако да представљају различите приступе решавању проблема нестабилности модела:

1. **OLS (Ordinary Least Squares)**
   Користи се као основни модел и референтна тачка за поређење.
2. **Ridge регресија**
   Уводи регуларизацију која смањује осетљивост модела на повезане атрибуте и доприноси стабилнијим параметрима.
3. **Lasso регресија**
   Омогућава поједностављење модела кроз елиминацију мање значајних променљивих, што утиче на интерпретабилност.
4. **Elastic Net регресија**
   Комбинује особине Ridge и Lasso регресије и користи се у случајевима када постоји више међусобно повезаних атрибута.

5. **Huber регресија**
   Представља робусни приступ који смањује утицај екстремних вредности, при чему задржава структуру линеарног модела.

Избор ових метода омогућава анализу разлике између класичног, регуларизованог и робусног приступа линеарној регресији.

---

4. Начин евалуације

Квалитет модела биће оцењиван коришћењем следећих метрика:

- Root Mean Squared Error (RMSE)
- Прилагођени R^2 коефицијент детерминације

Поред тога, анализираће се и стабилност регресионих коефицијената кроз више експеримената, као показатељ поузданости и интерпретабилности модела.

---

5. Технологије

За реализацију пројекта користиће се:

- Python 3.12

Библиотеке:

- NumPy
- Pandas
- Matplotlib
- Scikit-learn

---

6. Примери сличних готових решења

- Анализа и визуелизација података о ценама некретнина уз примену регресионих модела https://www.kaggle.com/code/pmarcelino/comprehensive-data-exploration-with-python
- Примена више регресионих метода и њихово комбиновање ради побољшања тачности предвиђања https://www.kaggle.com/code/serigne/stacked-regressions-top-4-on-leaderboard

---

7. Литература

- https://scikit-learn.org/stable/modules/linear_model.html
- https://en.wikipedia.org/wiki/Linear_regression
- https://en.wikipedia.org/wiki/Ridge_regression

- https://en.wikipedia.org/wiki/Lasso_(statistics)
- https://en.wikipedia.org/wiki/Huber_loss