

CBU Kodlash Tanlovi: Ma'lumotlar tahlili va qayta ishlash

Muallif: Miradxamov Mirafzal

Sana: 2024-yil 15-avgust



Kirish

Ushbu taqdimot moliyaviy ma'lumotlar to'plamlari (dataset1.csv, dataset2.csv, dataset3.csv) ustida ma'lumotlarni qayta ishlash va tahlil qilishga bag'ishlangan.

1

Ma'lumotlarni yuklash

Kutubxonalarni import qilish va ma'lumotlarni yuklash.

2

Ma'lumotlarni tozalash

Ustunlarni qayta nomlash va anomaliyalarni bartaraf etish.

3

Birlashtirish va to'ldirish

Ma'lumotlar to'plamlarini birlashtirish va yetishmayotgan qiymatlarni to'ldirish.

4

Metrikalar va vizualizatsiyalar

Xato metrikalarini hisoblash va natijalarni vizuallashtirish.

Kutubxonalar va Ma'lumotlarni Yuklash

Tahlil uchun zarur bo'lgan asosiy kutubxonalar import qilindi:

- Pandas (ma'lumotlar manipulyatsiyasi)
- Numpy (raqamli hisob-kitoblar)
- Matplotlib & Seaborn (vizualizatsiya)
- Scikit-learn (mashinani o'rganish)



Ma'lumotlar to'plamlari yuklandi va dastlabki o'rGANISH amalga oshirildi (df.head(), df.info(), df.describe()).



Ma'lumotlarni Tozalash: Ustunlarni Qayta Nomlash

Ma'lumotlar to'plamlarida ustun nomlarining izchilligini ta'minlash uchun qayta nomlash operatsiyalari bajarildi.



'Date' ustuni

Barcha ma'lumotlar
to'plamlarida 'date' deb
nomlandi.



'ID' ustuni

Barcha ma'lumotlar
to'plamlarida 'id' deb nomlandi.



Izchilllik

Birlashtirish va tahlil qilishni
osonlashtirish.

Ma'lumotlarni Tozalash: Anomaliyalarni Bartaraf Etish

'actual_value' va 'predicted_value' ustunlaridagi anomaliyalar IQR (Interquartile Range) usuli yordamida aniqlandi va olib tashlandi.

- **Q1 va Q3 hisoblash:** Ma'lumotlarning 25% va 75% kvartillari.
- **IQR hisoblash:** $Q_3 - Q_1$.
- **Chegaralarni aniqlash:** $Q_1 - 1.5 \cdot IQR$ va $Q_3 + 1.5 \cdot IQR$.
- **Anomaliyalarni olib tashlash:** Chegaradan tashqaridagi qiymatlar.



Ma'lumotlar To'plamlarini Birlashtirish

Uchta ma'lumotlar to'plami ('dataset1', 'dataset2', 'dataset3') 'id' va 'date' ustunlari bo'yicha birlashtirildi.



Dataset1

Asosiy ma'lumotlar to'plami.

Dataset2

Qo'shimcha ma'lumotlar.

Dataset3

Uchinchi ma'lumotlar manbai.

Birlashtirilgan DF

pd.merge() yordamida 'inner' usulida.

Yetishmayotgan Qiymatlarni To'ldirish

Ma'lumotlar to'plamidagi yetishmayotgan qiymatlar (NaN) tegishli usullar bilan to'ldirildi.

Raqamli ustunlar

O'rtacha (mean) qiymat bilan to'ldirildi.

Kategorik ustunlar

Eng ko'p takrorlanadigan (mode) qiymat bilan to'ldirildi.

Maqsad

Ma'lumotlar sifatini oshirish va modelni o'qitishga tayyorlash.

Metrikalarни Hisoblash

'actual_value' va 'predicted_value' о'rtasidagi xato metrikalari (MAE va RMSE) hisoblandi.

99.75% 95.15% 100.00.. 97.51%

Aniqlik

Modelning umumiy
aniqligi.

Precision

Defolt sinfi uchun aniqlik.

Recall

Defolt sinfi uchun to'liqlik.

F1-score

Precision va Recall
o'rtasidagi muvozanat.

1.0000

AUC

ROC egri chizig'i ostidagi
maydon.

Vizualizatsiyalar

Ma'lumotlar to'plamini chuqurroq tushunish uchun turli xil vizualizatsiyalar yaratildi.

Bu grafiklar ma'lumotlar taqsimoti, anomaliyalar va ustunlar orasidagi bog'liqliklarni ko'rsatadi.

Xulosa va Manbalar

Ushbu loyiha moliyaviy ma'lumotlarni tahlil qilish va qayta ishlashning to'liq jarayonini namoyish etdi.

Asosiy xulosalar:

- Tozalangan ma'lumotlar to'plami modelni o'qitishga tayyor.
- Yuqori aniqlikdagi xato metrikalari modelning samaradorligini ko'rsatadi.
- Vizualizatsiyalar ma'lumotlar haqida qimmatli tushunchalar beradi.

Manbalar:

- Jupyter Notebook
- Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn kutubxonalari