Domácí úkol: Lineární regrese pro předpověď šance na přijetí na vysokou školu

Odevzdání

- Úkol odevzdejte do 12.5.2024 do 23:59 hodin.
- Odevzdejte zdrojový kód v Jupyter notebooku, který obsahuje všechny kroky od načítání dat po evaluaci modelu.
- V Jupyter notebooku přiložte krátké popisky nebo komentáře popisující vaše postupy, zjištění a interpretace výsledků.

Cíl

Vytvořte model lineární regrese, který na základě dostupných údajů předpoví šanci na přijetí na vysokou školu.

Dataset

Váš dataset obsahuje následující proměnné:

- GRE skóre (maximálně 340 bodů)
- TOEFL skóre (maximálně 120 bodů)
- Hodnocení univerzity (1 až 5 bodů)
- Síla motivačního dopisu a doporučení (1 až 5 bodů)
- Průměrná známka z bakalářského studia (maximálně 10 bodů)
- Zkušenosti s výzkumem (buď 0 nebo 1)
- Šance na přijetí (hodnota mezi 0 a 1)

Úkoly

Příprava dat

1. Načtěte dataset a ujistěte se, že rozumíte hodnotám ve sloupcích. Můžete změnit názvy sloupců a odstranit z nich mezery, což se vám může později hodit.

Normálnost dat pro TOEFL skóre

1. Vypočítejte průměr a medián pro TOEFL skóre.
2. Vypočítejte směrodatnou odchylku.
3. Vytvořte histogram pro TOEFL skóre a nezapomeňte zobrazit odhadovanou křivku pravděpodobnostního rozdělení.
4. Otestujte, zda je TOEFL skóre vybráno z normálního rozdělení pomocí statistického testu. Uvažujte hladinu.

Testování hypotéz

1. Rozdělte data na studenty, kteří mají a nemají zkušenost s výzkumem.
2. Pro obě skupiny spočítejte průměrné GRE skóre.
3. Statisticky otestujte, zda existuje statisticky významný rozdíl v GRE skóre mezi těmito dvěma skupinami studentů. Vyberte správný test a nezapomeňte krátce okomentovat vyhodnocení testu.

Korelace

1. Vypočítejte korelaci mezi TOEFL skóre a GRE skóre.
2. Krátce okomentujte, co znamená vypočítaná korelace. Je tato korelace vysoká, pozitivní/negativní?
3. Vytvořte korelační graf (sns.regplot) pro vysvětlovanou proměnnou šance na přijetí (Chance of Admit) a proměnnou, která je s ní nejvíce korelovaná.

Vysvětlovaná proměnná

1. Naší vysvětlovanou proměnnou bude šance na přijetí (Chance of Admit). Vytvořte boxplot pro tuto proměnnou, aby bylo možné vizuálně identifikovat případné odlehlé hodnoty.
2. Odstraňte odlehlé pozorování na základě kritérií zjištěných z boxplotu.

Lineární regrese

1. Sestavte rovnici pro lineární regresi. Do rovnice zahrňte všechny proměnné, které dávají smysl. Pozor, možná bude potřeba názvy sloupců přejmenovat.
2. Odhadněte parametry lineárního regresního modelu.
3. Interpretujte koeficienty modelu. Které koeficienty jsou statisticky významné? Diskutujte, které proměnné mají největší vliv na šance na přijetí a proč.
4. Vyhodnoťte kvalitu fitu vašeho lineárního regresního modelu pomocí koeficientu determinace R^2.
5. Spočítejte Cookovu vzdálenost pro jednotlivé body.
6. Vytvořte histogram pro Cookovu vzdálenost. Je potřeba nějaké body odstranit, protože by moc ovlivňovaly naší lineární regresi?

Bonus: Cookova vzdalenost

1. Spočítejte Cookovu vzdálenost pro jednotlivé body.
2. Vytvořte histogram pro Cookovu vzdálenost. Je potřeba nějaké body odstranit, protože by moc ovlivňovaly naší lineární regresi?