Věrohodnost

[Celkem 4 body] V souboru Data_2024.xlsx (v listu Data_věrohodnost) máte zaznamenáno, jak dlouho po ukončení vzdělání pracují absolventi VUT ve „svém“ oboru (v letech). Někteří absolventi však po nějaké (zaznamenané) době přestali reagovat. U těchto absolventů je znám čas kdy přerušili kontakt (ale kdy ještě pracovali v oboru). Tato pozorování berte jako zprava cenzorovaná (jsou označena ve sloupečku „cesored“ jedničkou). Předpokládejte, že doba zaměstnání v oboru se řídí Weibullovým rozdělením pravděpodobnosti začínajícím v 0
(parametr prahu-threshold nastavte na 0).

1) Zapište zvolenou parametrizaci Weibullova rozdělení, logaritmickou-věrohodnostní funkci pro zadaná data a její parciální derivace podle parametrů (shape, scale).
2) Pomocí scipy.optimize nalezněte maximálně věrohodné odhady parametrů weibullova rozdělení.
3) Pomocí věrohodnostního poměru otestujte hypotézu, že exponenciální rozdělení je postačujícím modelem zapsaných dat (Parametr tvaru = 1)
4) Podle výsledku ze 3) použijte výsledné rozdělení pravděpodobnosti (s maximálně věrohodnými odhady jako parametry) a nalezněte bodové odhady pro střední dobu zaměstnání v oboru a 10% percentil zaměstnání v oboru (za jakou dobu odejde do jiného odboru 10 % absolventů).
5) [dobrovolná část] zkuste nějak slovně charakterizovat/popsat fungování doby zaměstnání v oboru jako náhodné veličiny, dle Vašich výsledků a parametrů

Regrese – MSP projekt zadání

[celkem 8 bodů]
Disclaimer: data (včetně „příběhu“) jsou vygenerovaná a nemusí mít dobrý obraz v realitě. Berte, proto prosím výsledky z regrese s „rezervou“. Díky.

Podařilo se Vám pomocí stroje času vrátit do doby „zlatého věku“ sociálních sítí a rozhodli jste se konkurovat Facebooku a Twitteru. V souboru Data_2024.xlsx (v listu Data_regrese) máte k dispozici záznamy od více než 500 uživatelů o rychlosti odezvy (sloupec ping [ms]) během používání Vaší aplikace. Ke každému zápisu máte navíc k dispozici údaje o počtu uživatelů (sloupec ActiveUsers) v daném okamžiku, o procentu uživatelů, kteří momentálně interagují s prezentovaným obsahem (sloupec InteractingPct), o procentu uživatelů, kteří jen tupě scrollují po Vaší obdobě timeline/twitterfeedu (sloupec ScrollingPct) a o operačním systému zařízení ze kterého se uživatel připojil (OSType).

1) Pomocí zpětné eliminace určete vhodný regresní model. Za výchozí „plný“ model považujte plný kvadratický model (všechny interakce druhého řádu a všechny druhé mocniny, které dávají smysl).
    - Zapište rovnici Vašeho finálního modelu.
    - Diskutujte splnění předpokladů lineární regrese a základní regresní diagnostiky.
    - Pokud (až během regresního modelování) identifikujete některé „extrémně odlehlé hodnoty“ můžete ty „nejodlehlejší“ hodnoty, po alespoň krátkém zdůvodnění, vyřadit.

2) Pomocí Vašeho výsledného modelu identifikujte, pro které nastavení parametrů má odezva nejproblematičtější (největší) hodnotu (použijte model, nikoli samotná pozorování).
3) Odhadněte hodnotu odezvy uživatele s Windows, při průměrném nastavení ostatních parametrů a vypočtěte konfidenční interval a predikční interval pro toto nastavení.
4) Na základě jakýchkoli vypočtených charakteristik argumentujte, zdali je Váš model „vhodný“ pro další použití.