

Analýza leteckých nehod

Anastasiia Osadchuk

1 Situace

Letecké havárie jsou klíčovým bezpečnostním a veřejným zájmem. Analýza těchto událostí může odhalit hlavní příčiny a pomoci v prevenci. Rozhodla jsem se analyzovat data o leteckých nehodách, abych identifikovala příčiny, počet obětí a časové rozložení těchto událostí.

2 Úkol

Analyzovat data o leteckých haváriích, vizualizovat je pomocí Pythonu a vytvořit model pro predikci počtu havárií v následujících 10 letech pomocí ARIMA modelu.

3 Postup

Data o leteckých haváriích byla stažena z databáze Kaggle. Obsahovala informace o datu, místě, provozovateli, typu letadla, počtu obětí, přeživších atd.

Pro zpracování dat byl použit skript *process_data.py*, který:

- Vyčistil data od duplicit a nahradil chybějící hodnoty hodnotou NA.
- Odstranil sloupce, které nebyly použity v analýze.
- Přidal nové sloupce, jako například:
 - season (roční období)
 - total_fatalities (celkový počet obětí)

Data byla následně analyzována a vizualizována pomocí Pythonu. Provedla jsem několik predikcí před analýzou:

- Díky technologickému pokroku je nyní méně nehod
- Vrchol havárií byl během druhé světové války
- Více havárií na podzim a v zimě kvůli povětrnostním podmínkám

Vizualizovala jsem data, abych tyto předpoklady podložila.

Pro predikci budoucích havárií jsem použila model ARIMA. Model ARIMA je vhodný pro data, která ukazují určitou úroveň stacionarity po diferenciaci, což byl případ u našich ročních počtů havárií.

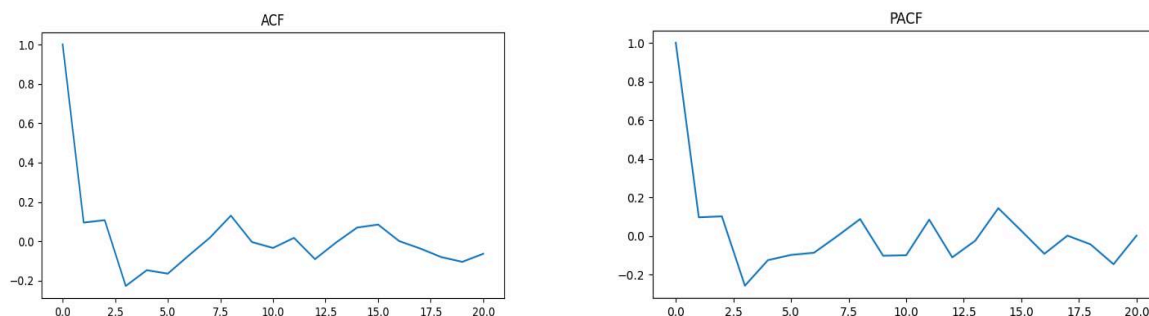
- Model ARIMA má tři parametry: p , d a q , které představují počet autoregresivních členů (AR), stupeň diferenciaci (I) a počet klouzavých průměrů (MA).

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad \text{nebo}$$

$$\begin{array}{ccccc}
 (1 - \phi_1 B - \dots - \phi_p B^p) & (1 - B)^d y_t & = & c + (1 + \theta_1 B + \dots + \theta_q B^q) \varepsilon_t \\
 \uparrow & \uparrow & & \uparrow \\
 \text{AR}(p) & d \text{ differences} & & \text{MA}(q)
 \end{array}$$

- Analyzovala jsem autokorelační funkci (ACF) a parciální autokorelační funkci (PACF), abych určila vhodné hodnoty parametrů p , d a q .

Grafy ACF a PACF jsou uvedeny níže.



Na základě analýzy grafů ACF a PACF jsem zvolila následující hodnoty parametrů:

- $d = 1$: Rychlý pokles v ACF grafu po prvním lagu naznačuje, že jedna diferenciací je dostačující k dosažení stacionarity.
- $p = 1$: Významný spike na prvním lagu v PACF grafu naznačuje, že jeden 1 lag je vhodný pro autoregresivní část modelu.
- $q = 1$: ACF graf ukazuje určitou úroveň autokorelace po prvním lagu, což naznačuje, že jeden lag v klouzavém průměru je vhodný.

Model byl natrénován na historických datech o ročních počtech havárií a poté byl použit k predikci počtu havárií na následujících 10 let.

4 Výsledky

Na základě mé analýzy jsem dospěla k následujícím závěrům:

- Počet havárií každý rok klesá, což může být důsledkem technologického pokroku a zvýšených bezpečnostních opatření.
- Nejvíce havárií se vyskytuje v zimě a létě, ale rozdíl není významný. Povětrnostní podmínky hrají roli, ale nejsou jediným faktorem.
- Většina havárií a úmrtí je způsobena lidskou chybou, což ukazuje na potřebu zlepšení školení a postupů pro piloty a letecký personál.
- Havárie způsobené povětrnostními podmínkami jsou nejčastější v létě a zimě, kdy jsou podmínky často extrémnější.
- Vrchol havárií byl během druhé světové války, což bylo způsobeno vysokým objemem vojenských operací a omezenými bezpečnostními protokoly.
- Většina společností a letadel, které zažily havárie, byly vojenské, což odráží intenzivní využívání vojenských letadel během konfliktů.

5 Reference

- Kaggle. (2023). Airplane Crashes and Fatalities Since 1908. Retrieved from <https://www.kaggle.com/datasets/saurograndi/airplane-cras>
- McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media. Montgomery, D. C., Peck, E. A., Vining, G.
- G. (2015). Introduction to Linear Regression Analysis. Wiley
- Nau, R. (n.d.). Introduction to ARIMA: nonseasonal models. Retrieved from <https://people.duke.edu/~rnau/411arim.htm>