## 2 – batch processing i analiza szeregu czasowego

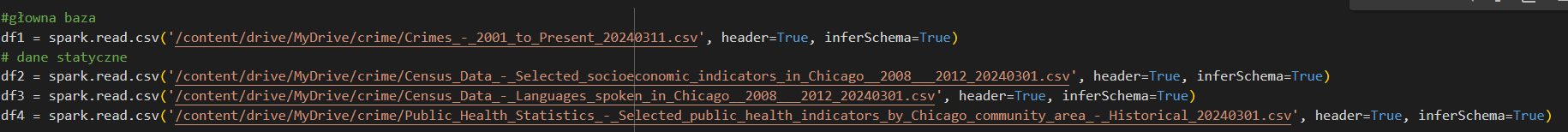
1. Processing danych:

* Załadowanie wszystkich niezbędnych bibliotek i otworzenie sesji sparkowej.

Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

* Załadowanie plików do sparka



Źródła danych statycznych do analizy socjotechnicznej:

<https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu/about_data>

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2/about_data>

https://data.cityofchicago.org/Health-Human-Services/Census-Data-Languages-spoken-in-Chicago-2008-2012/a2fk-ec6q/about\_data

* Czyścenie danych i joinny po kolumnie Community Area oraz archiwizacja na S3 w formacie parquet.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Uwagi:

Ze względu na bardzo dużą liczbę wartości null w kolumnach dotyczących współrzędnych i Community Area(po której są robione jony), pierwsze 2 lata utraciły by 90% zdarzeń przez co mieszały by tylko w analizach i predykcjach. Dlatego odrzucamy je od razu. W reszcie roczników jakiekolwiek nulle nie przekraczały więcej jak kilku tysięcy rekordów na rok, więc zwykły drop nie zaszkodzi. Wyciągamy daty i sortujemy przed jonowaniem bo wtedy robimy to na mniejszym dataframie.

1. Klasa PredictionModel

Agregacja danych do trzech szeregów czasowych, przejście z Sparka na Pandasa i użycie algorytmu Holt-Winters do predykcji dla każdego z tych szeregów, wizualizacja i ewaluacja modelu.

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Uwagi:

Trzy stopnie agregacji: dzienna, tygodniowa i miesięczna.

Każda z agregacji jest przycinana o ostatnią aktualizacje danych ze strony(aktualizowane co tydzień) w celu upewnienia się że dane z ostatniego okresu są załadowane w całości.

Dodatkowo agregacja tygodniowa zlicza tygodnie licząc od poniedziałku.

Ze względu na fakt że spark ML nie ma żadnych gotowych rozwiązań z zakresu predykcji szeregów czasowych zarówno pod względem indeksowania jak i samych modeli przechodzimy na Pandas i rozwiązania z biblioteki statsmodels.

Wybór algorytmu podyktowany jest dystrybucją danych, ładną sezonowością i wyraźnym trendem, a także minimalna liczbą parametrów do strojenia, przy jednocześnie akceptowalnych błędach w predykcji.

Do ewaluacji modelu mamy 3 metryki:

RMSE mocno każąca duże błędy nie odporna niestety na wartości odstające

MAE Jako uzupełnienie RMSE odporniejsza na wartości odstające

MAPE jako najbardziej intuicyjna podająca błędy pozwalająca porównać błędy pomiędzy różnymi formami agregacji.

1. Wywołanie funkcji dla 3 stopni agregacji, wizualizacja, raport z błędów oraz archiwizacja w postaci pliku CSV

Agregacja dzienna

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie

Tygodniowa

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznieMiesięczna

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie

Do zrobienia:

* Prawdziwa predykcja do przodu, czyli poza istniejące dane, jest już prototyp, wymaga jeszcze drobnych poprawek i umieszczenia wewnątrz klasy PredictionModel.

Ciężko było znaleźć wskazówki jak to zrobić, bo wszystkie szkolenia pokazują jedynie ewaluacje na zbiorze testowym gdzie widać jak działa algorytm, ale nie jest to predykcja w przyszłość. Wymaga to stworzenia dynamicznej listy dat, która będzie zaczynała się dzień po ostatnich zarejestrowanych danych, a kończyła na maksymalnej odległości na jaką chcemy przewidywać. Następnie cały zbiór danych używamy do treningu algorytmu, bez ewaluacji za pomocą metryk, a jedynie wizualizacji i zapisaniu wyników.

* Zwalczenie problemu „53 tygodnia” poprzez arbitralną decyzję czy usuwamy te dane, czy też dodajemy lub rozpraszamy po innych tygodniach.
* Czytelniejsza wizualizacja w bibliotece Ploty wraz naniesieniem prawidłowych metryk(odcinków czasowych) na skalę.
* Spięcie tego wszystkiego z S3 bucketami 😉