|  |  |  |  |
| --- | --- | --- | --- |
|  | Politechnika Bydgoska im. J. J. Śniadeckich  Wydział Telekomunikacji,  Informatyki i Elektrotechniki  **Zakład Systemów Teleinformatycznych** | |  |
| **Przedmiot** | ARiSC | | |
| **Prowadzący** | prof. dr hab. inż. prof. PBŚ Piotr Cofta | | |
| **Temat** | *Project Raport 1* | | |
| **Student** | Cezary Tytko | | |
| **Ocena** |  | **Data oddania spr.** |  |

Etap 1. – wstępna analiza danych

Zacząłem od wczytania danych (dla tygodnia 37 i 38), wczytanie nie spowodowało żadnych błędów, zatem zakładam że jak na razie są one przynajmniej poprawnie sformatowane, zakres czasów też się zgadza, jak i liczność pomiarów jest zgodna z oczekiwaną czyli 604800 (7 \* 24 \* 60 \* 60), danych nam nie brakuje, ani nie jest ich nadto, należy jeszcze sprawdzić czy są one odpowiednio rozmieszczone.

Aby sprawdzić rozmieszczenie, obliczyłem różnicę kolejnych czasów, wiedziałem że wartości te nie będą idealnie równe, ale powinny być bardzo blisko 1s i w istocie tak było, dla pierwszego tygodnia największy odstęp jaki odczytałem wyniósł 1.174 ms, a najmniejszy (należałoby się spodziewać że średnia z tych dwóch wartości powinna być w okolicy 1s) wyniósł 814 ms, tak odległych wartości od zakładanej było bardzo niewiele, a i tak uważam że taki odstęp jest do przyjęcia, zatem dane dla pierwszego tygodnia są poprawne jeżeli chodzi o ich rozmieszczenie w czasie. Dla drugiego tygodnia największa różnica wynosi 2.230 ms a najmniejsza 123ms, jednak jest to pojedynczy przypadek, i pomimo odstępu ponad 2 sekundy liczność danych się zgadza oraz taka sama anomalia występuje w tym samym momencie dla wszystkich morzonych parametrów, zakładam (nie wiem tego na pewno) że mogło być spowodowane opóźnieniem programu/bazy danych, a pomiary zostały zakolejkowane, dlatego po długiej przerwie kolejne zapisy następowały z przyśpieszeniem i aż wyrównały do interwału 1 s

Aby dostosować dane do przyszłej analizy zdecydowałem zresetować index czasu i założyć stały interwał 1 s dla wszystkich pomiarów, aby wyeliminować wartości z poziomu ms, oraz naprawić rozłożenie danych gdzie przerwa wynosiła 2s, a następnie 0,1 s. W takim przypadku zwykłe zaokrąglenie, od 0,5 , w górę, czy w dół nie zadziała nie tylko ze względu na różnicę ponad 2 s, ale nawet przy prawie stałym interwale 1 s możemy znaleźć się w punkcie gdzie z naszej perspektywy nieistotna różnica sprawi że w wyniku zaokrąglenia dwa pomiary otrzymają identyczny czas, oraz w parze do tego powstanie gdzieś luka, dlatego lepiej wygenerować sztucznie indeksy i je przypisać w kolejności jakiej dane były zapisane.

Podstawowe parametry:

1. Tydzień

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Wyliczone parametry zgadzają się s tym czego należało się spodziewać, czyli napięcia w okolicy 230 V z pewnymi odchyłami, oraz średniej częstotliwości 50 Hz z niewielki odstępstwami, thdv1 niewiele mówi nam w prostu, ale może okazać się przydatne w późniejszym etapie, związanym z uczeniem i przewidywaniem wartości.

1. Tydzień

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Analogicznie jak dla pierwszego tygodnia dane wyglądają dobrze i nie ma anomalii w postaci np. zaniku napięcia

Histogramy:

Obraz zawierający tekst, zrzut ekranu, diagram

Opis wygenerowany automatycznie

Postanowiłem nałożyć na wykresy hipotetyczny rozkład normalny dla takich danych i wizualnie sprawdzić czy się one pokrywają w znacznym stopniu (dla v1 od razu można stwierdzić że rozkład ten nie jest normalny).

Obraz zawierający zrzut ekranu, diagram, Wykres, tekst

Opis wygenerowany automatycznie

Widać że zdecydowanie wykres thdv1 pokrywa się z rozkładem normalny, to samo można by powiedzieć o wykresie dla f1, z wyjątkiem zdecydowanie wyróżniającej się wartości bardzo bliskiej 50 HZ, która to jest oczekiwaną i zadana wartością w sieci elektrycznej, natomiast dla wykresu v1 widzę połączenie trzech rozkładów normalny, na tym etapie niestety niewiele więcej jestem wstanie powiedzieć, nie wiem jak dokładnie wyglądały warunki badania, mógłbym np. przypuszczać że mamy do czynienia z trzema różnymi stanami w sieci, które mogą być określane przez lokalne obciążenie sieci urządzeniami znacznie oddziaływać na napięcie, jak silniki eklektyczne, kondensatory, grzałki.

Obraz zawierający diagram, Wykres, zrzut ekranu, tekst

Opis wygenerowany automatycznie

Takie same wnioski wyciągnąłbym z histogramów dla drugiego tygodnia, jak dla poprzedniego

Etap 2 – przewidywanie wartości

Etap ten zacząłem od standaryzacji danych i na takich danych będę opierał się w dalszej części raportu.

Na początek sprawdziłem korelację i auto korelację(z różnymi wartościami lag):

Obraz zawierający tekst, zrzut ekranu, menu, Czcionka

Opis wygenerowany automatycznie

Jak widać występuje raczej niewielka korelacja miedzy kolumnami, największa, ale wynosząca tylko -0,12 do 0,147 w zależności od sposobu obliczania, dla pary thdv1 i v1, znak minus mówi na tylko tyle że korelacja jest odwrotnie proporcjonalna, ale i tak jest największa z pośród zaprezentowanych par. Autokorelacja dla f1 i v1 jest bardzo wysoka i stopniowo, ale powoli spada wraz ze zwiększaniem odstępu, natomiast dla thdv1 jest mniejsza wynosi 0,18 dla wartości po jednej sekundzie i zdecydowanie spada dla większych odstępów.

Sprawdziłem jeszcze widno częstotliwości stosując transformatę Fouriera per kolumna, otrzymałem takie wykresy:

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznie

Dla v1 nie wykryto innych częstotliwości, dla thdv1 mamy niewielkie widmo w całym zakresie, możemy tutaj mówić o szumie, dla f1 widmo częstotliwości skupia się przy wartości 0, z pewnymi odchyleniami, przypominającymi rozkład normalny, oraz widzimy dwie pary częstszych wartości niż szum.

Można by zastosować teraz dodatkowe techniki analizy danych, jak uśrednianie, interpolację, analizę harmonicznych, ale zdecyduję się pozostawić dane jedynie ustandaryzowane i zastosować techniki głębokiego uczenia w których nie wymaga się dodatkowej analizy, ponieważ model ma zrobić to za nas.

Wykorzystam model posiadający na wejściu warstwę rekurencyjną, np. LSTM, albo RNN, ponieważ są to warstwy stworzone do analizy szeregów czasowych z którym mamy do czynienia w zadaniu, neurony takiej warstwy poza przyjmowaniem wartości cech, przyjmują również jako parametry wartości wyjść z poprzedniej iteracji. Dodatkowo przygotowałem funkcję to tworzenia kolejnych szeregów czasowych, o zadanej długości i wskazanej odległości do wartości przewidywanej. Jako funkcję straty wykorzystam MSE, a dodatkowo przy walidacji modelu będę opierał się ma mierze R2 z takimi narzędziami. Z takim zestawem narzędzi można przystąpić to modelowania sieci i jej treningu.