Analiza zależności średniego spalania w transporcie drogowym od warunków pogodowych

Kacper Dusza
Nr indeksu: 266868

Czerwiec 5, 2023

Spis treści

1 Wprowadzenie do problemu

Transport drogowy zarówno w Polsce, jak i na całym świecie jest jednym z głównych środków do przemieszczania towarów. Przewóz lądowy jest szczególnie skuteczny w przypadku dystrybucji towarów na krótkie i średnie odległości. Samochody ciężarowe są powszechnie stosowane do dostarczania towarów w ramach miast, regionów czy między sąsiednimi krajami. Mając to na uwadze postanowiłem zbadać, zależność spalania od warunków pogodowych.

2 Cel projektu

Celem projektu jest próba przewidywania średniego spalania samochodu ciężarowego na podstawie warunków pogodowych.

3 Pozyskiwanie danych

3.1 Dane dotyczące tras

Dane dotyczące odcinków przebytych przez ciężarówkę oraz spalania zostały pozyskane od firmy transportowej. Dokładniej pochodzą one z GPS zamontowanego w kabinie pojazdu, który zbiera informacje na temat lokalizacji rozpoczęcia i zakończenia trasy, spalania (poprzez podłączenie z komputerem pojazdu) i średniej prędkości. Poza tym jest również data oraz godzina odbytych tranzytów, długość trasy oraz stan licznika. Warto dodać, iż pozyskane dane są z przestrzeni 2,5 lat oraz dotyczą jednego samochodu ciężarowego z jednym pracownikiem. Dane zostały zebrane z urządzenia GPS do zbiorczego pliku CSV, którego przetwarzaniem następnie się zająłem.

Rekordy zawarte w pliku z trasami:

- data i godzina początku oraz końca trasy,
- lokalizacja początku i końca trasy,
- długość trasy,
- czas trasy,
- średnia prędkość,
- średnie zużycie paliwa,
- stan licznika.

3.2 Dane dotyczące pogody

Dane dotyczące pogody zostały pobrane poprzez API z strony `https://open-meteo.com/`, kod natomiast znajduje się w pliku "data_collecting_methods". Aby je pozyskać należało zamienić nazwę miasta w szerokość i długość geograficzną. Szczegółowy sposób pobierania przedstawię poniżej.

1. Posiadając dane początku trasy oraz czas jej trwania pobrane zostały dane pogodowe z całego okresu czasu dla miejsca rozpoczęcia trasy.
2. Posiadając dane zakończenia trasy oraz wcześniej wspomniany czas jej trwania pobrane zostały również dane z całego okresu czasu dla miejsca zakończenia trasy.
3. Następnie posiadając obie tabele w każdej z nich wyliczałem średnią dla poszczególnych danych.
4. Z dwóch tabel, w których jako klucz potraktowałem datę rozpoczęcia trasy złączyłem dane pogodowe również je uśredniając.

Przykładowo: Trasa z Wrocławia do Drezna rozpoczyna się o 12:30 i trwa 5 godzin. Na początku pobieram dane pogodowe dla Wrocławia z godzin: 13:00, 14:00, 15:00, 16:00, 17:00. Następnie dane te zostają uśrednione aby posiadać po jednej danej każdego rodzaju, przykładowo dla pięciu uśrednionych odczytów temperatury otrzymuję jeden, który jest zachowany w pliku. W dalszej kolejności robię to samo dla Drezna. Gdy posiadam dane zarówno dla Drezna jak i Wrocławia łącze obydwie tabele również uśredniając ich wartości dla obu miast, aby dane były maksymalnie zbliżone do warunków jakie panowały na danym odcinku drogi.

Do danych pogodowych należą:

- temperatura,
- wilgotność powietrza,
- prędkość wiatru,
- opady atmosferyczne.

4 Scalenie danych do jednej tabeli

Po pozyskaniu wszystkich danych rzeczą kluczową okazało się złączenie plików CSV w jeden, w celu ułatwienia dalszej analizy. Na tym etapie posiadałem 3 tabele:

- "weather_start_city tabela z warunkami pogodowymi w odniesieniu do miasta, z którego rozpoczynała się trasa,
- "weather_destination_city tabela z warunkami pogodowymi z miasta końcowego,
- "Odcinki tras dla pojazdu tabela z trasami pozyskanymi z GPS samochodu ciężarowego.

Atrybutem umożliwiającym połączenie był czas rozpoczęcia trasy. W pierwszej kolejności zostały połączone dane pogodowe, w wcześniej wspomniany sposób, to znaczy uśredniając dane z obydwu tabel dla konkretnej trasy. W dalszej kolejności dołączona została tabela z odcinkami tras.

5 Uzupełnianie brakujących danych

5.1 Brakujące dane w pliku z trasami

Pierwszym z napotkanych problemów z brakującymi danymi miał miejsce w pliku z odcinkami tras pojazdu. Niestety okazało się, że w przypadku miejsc rozpoczęcia lub zakończenia trasy w Niemczech brakowało dokładnej lokalizacji. Tylko w niewielu rekordach było podane miasto, a w pozostałych jedynie informacji w postaci: "dr. A2". Niestety informacje te były zbyt niedokładne aby wywnioskować z nich odpowiednie współrzędne geograficzne. Problem ten rozbiłem na sześć etapów.

1. Jako pierwszy z etapów zdefiniowałem słownik wraz z domyślnymi wartościami współrzędnych dla najpopularniejszych rekordów powiązanych z autostradami. Aby precyzyjniej określić miejsca które powinny odpowiadać wpisom dotyczącym samych numerów autostrad starałem się wydedukować, biorąc pod uwagę punkty początkowe gdzie w danym czasie powinien zatrzymać się tir. Było to możliwe ze względu na powtarzalność tras oraz stosunkowo nie dużą ilość rekordów.

2. Posiadając dane zakończenia trasy oraz wcześniej wspomniany czas jej trwania pobierałem dane z całego okresu czasu dla miejsca zakończenia trasy. Słownik ten możemy zaobserwować w funkcji o nazwie "find_highway_coordinates". Nie byłem jednak w stanie zapisać w nim wszystkich powtarzających się autostrad, a jedynie te ukazujące się najczęściej. W przypadku gdy nazwa danej autostrady nie znajdowała się w słowniku lokalizacja była szacowana przez moduł geopy.

3. W przypadku dróg innych niż autostrady (łatwo je było rozpoznać po tym że nie posiadały literki A przed numerem) zdałem się całkowicie na moduł geopy. Jednak zauważyłem, że aby podawał dokładniejsze lokalizacje należy przed nazwą drogi dodać literkę 'b', przykładowo b112.

4. W niewielu rekordach występowała nazwa miasta, którą wystarczyło wyciągnąć regexem i bez problemu były znajdowane współrzędne przez moduł geopy.

5. Jeżeli w lokalizacji była jedynie nazwa kraju moduł geopy pobierał współrzędne dla centralnej części, co było sporym przybliżeniem, ale uważam takie rozwiązanie za słuszne.

6. W bardzo niewielu przypadkach, kiedy moduł geopy nie mógł sobie poradzić z lokalizacją rekordy te były usuwane. Głównie za sprawą błędnego wpisu, który definiował stację benzynową jako miasto. Na szczęście było tylko kilka takich wpisów.

5.2 Brakujące dane w finalnej tabeli

W tabeli powstałej z połączonych wcześniej tabel znalazł się jedynie jeden nieprawidłowy rekord, posiadający zera w całym wierszu. Z uwagi na to, iż był to jedyny taki wpis uznałem, że najlepszym rozwiązaniem będzie jego usunięcie. W innym przypadku posiadał by on negatywny wpływ na analizę danych oraz zakłamywał by odczyty.

6 Wstępna analiza i usunięcie zbędnych danych

6.1 Wstępna analiza w tabli trasy

Pierwsza z wstępnych analiz danych dotyczyła tabeli z trasami. Już na poziomie generowania pliku CSV zaznaczyłem, aby trasy były nie krótsze niż 100km, co powinno korzystnie wpłynąć na zależności, gdyż krótkie odcinki przeważnie wiązałyby się z wysokim spalaniem, co mogło by zachwiać wyniki. W przypadku tej tabeli zbędne okazały się również niektóre z kolumn, szczegóły przedstawiam poniżej.

1. Pracownik - była to pusta kolumna, wygenerowana przez system, powód odrzucenia jest oczywisty.
2. Lokalizacja początku trasy - podczas łączenia tabel doszedłem do wniosku, że kolumna ta w dalszej części pracy będzie bezwartościowa, albowiem jej rola zakończyła się podczas pozyskiwania danych współrzędnych do API pogodowego.
3. Lokalizacja końca trasy - podobnie jak w punkcie powyżej.
4. Odcinki tras dla pojazdu - tabela z trasami pozyskanymi z GPS tira.
5. Czas trasy - Informacja ta była z uwagi na średnią prędkość, która znajdowała się już w tabeli.
6. Czas postoju - również zbędna kolumna, nie było dla niej zastosowania.
7. Maksymalna prędkość - podobnie jak w punkcie powyżej.
8. Zużyte paliwo - kolumna ta była nadmiarowa, z uwagi na średnie spalanie, które wystarczyło do modelu.
9. Notatka - pusta kolumna.
10. Inne parametry - również pusta kolumna.

6.2 Przedstawienie danych i ich typów

W finalnej tabeli (weather_route_data) znalazło się 525 wierszy danych, a ich typy to:

- start_time - obiekty typu datetime, reprezentujący początek trasy
- end_time - obiekty typu datetime, reprezentujący koniec trasy
- temperature - liczba podwójnej precyzji, reprezentuje dane o temperaturze zewnętrznej podczas odbytej trasy
- humidity - liczba podwójnej precyzji, zawierająca dane o wilgotności powietrza
- wind_Speed - liczba podwójnej precyzji, z danymi o prędkości wiatru, niestety nie jest określony kierunek wiatru
- precipitation - liczba podwójnej precyzji informująca o opadach w milimetrach słupa wody na metr kwadratowy
- route_length - liczba podwójnej precyzji określająca długość trasy
- average_speed - liczba podwójnej precyzji określająca średnią prędkość
- ave_fuel_cons_per_100km - liczba podwójnej precyzji, określa średnie spalanie w litrach na 100 km
- speedometer - liczba podwójnej precyzji zawierająca informacje o stanie licznika

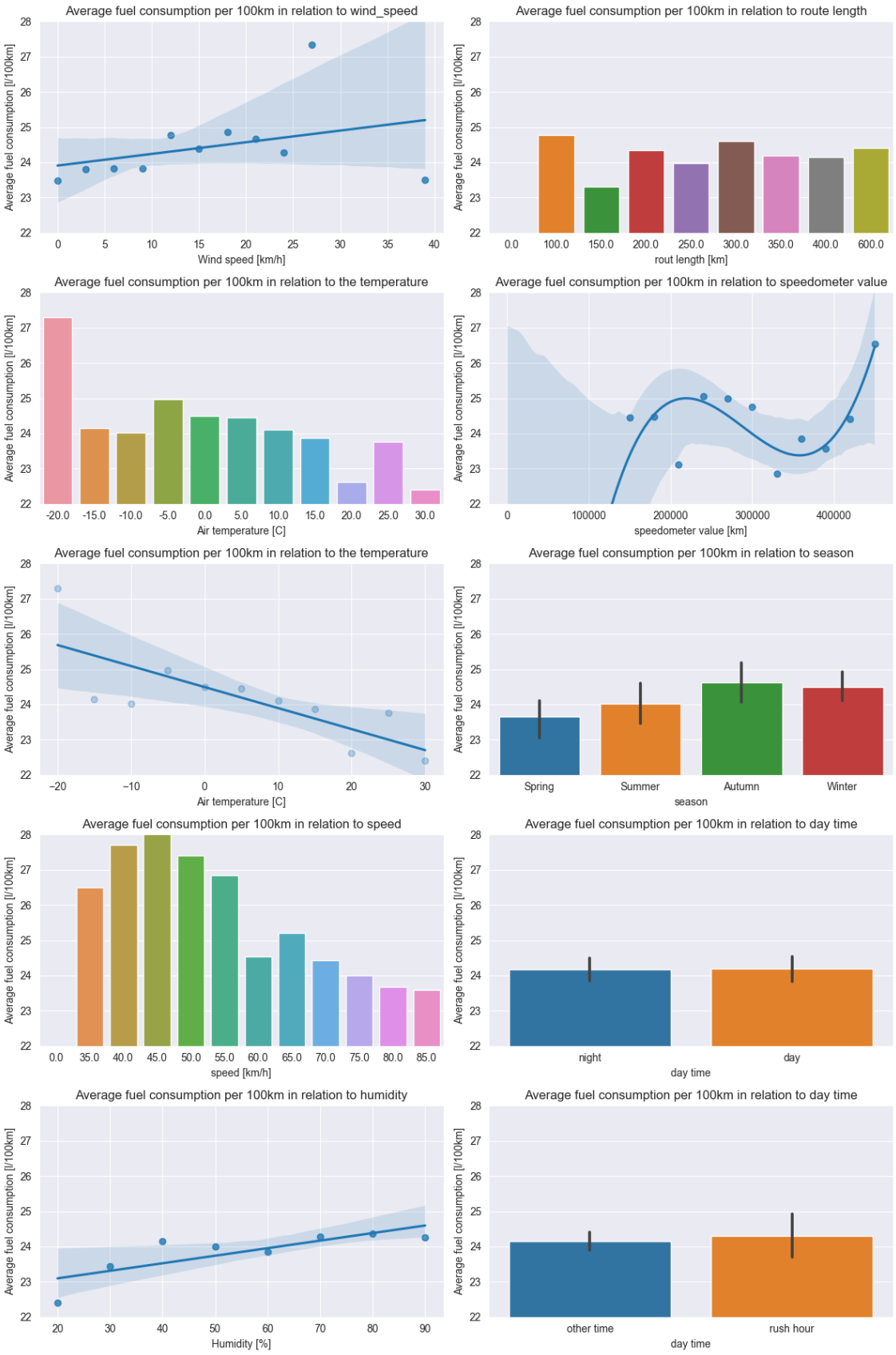
Średnia temperatura z wszystkich odczytów wynosi 9.5°C, co jest zgodne z średnimi danymi pogodowymi za te lata. Większość tras natomiast odbywała się z średnią prędkością 88.6 km/hm czyli prędkością, gdzie spalanie powinno być optymalne i jest ono w większości równe 25.8l/100km.

6.3 Wstępna analiza w głównym pliku Jupyter Notebook

Wstępna analiza w pliku Jupyter Notebook (route_analyzing.ipynb) polegała na stworzeniu tabel zależności danych od średniego spalania. Z powodu dużego odchylenia danych do wstępnej analizy niektóre z nich postanowiłem pogrupować, abym mógł zaobserwować powiązania. Dzięki takiemu podejściu już na wstępie mogłem zaobserwować, że dane o opadach są zbyt nieliczne aby mogły sprawdzić się w modelu. W kolumnie z opadami występowały w przeważającej części zera, w rezultacie czego postanowiłem ominąć tę kolumnę w dalszej analizie. Reszta danych w zaobserwowanych tabelach wydawała się zasadna.

6.4 Dalsza analiza na podstawie wykresów

Do analizy w tej części stworzyłem wykresy z pomocą bibliotek seaborn oraz matplotlib. W niektórych przypadkach aby wykresy były zwięźlejsze również użyłem wcześniej pogrupowanych danych. Do zobrazowania zależności użyłem wykresów słupkowych oraz punktowych z linią regresji. Dla lepszej czytelności i możliwości porównania stopnia zależności pomiędzy danymi postanowiłem umieścić wszystkie wykresy na kolejnej stronie.



Rysunek 1: Wykresy zależności danych od spalania

Pierwszy z wykresów przedstawia **zależność spalania od prędkości wiatru**. Z uwagi na różne kierunki wiatru postanowiłem posłużyć się wykresem punktowym z regresją aby zaobserwować ogólny trend. Różnica w spalaniu jest zauważalna na korzyść niskiej prędkości wiatru, co jest zgodne z założeniami. Jak możemy zaobserwować pomimo pominięcia cechy kierunku wiatru nadal jesteśmy w stanie wydedukować zależność pomiędzy tymi dwiema zamiennymi. Na tym etapie uważam, że cecha prędkość wiatru powinna znaleźć się jako jedna z składowych modelu.

Dwa kolejne wykresy przedstawiają **związek spalania od temperatury powietrza**. Dla lepszego zobrazowania w tym przypadku posłużyłem się obydwoma typami wykresów. Zależność ta wydaje się stosunkowo silna, albowiem różnica w spalaniu w przypadku amplitudy temperatur sięgającej 50 stopni jest na poziomie nawet około 5 litrów.

Następny z wykresów to **zależności spalania od prędkości pojazdu**. Ten przypadek wydawać by się mógł niezgodny z logiką, która początkowo mogłaby podpowiadać, że wraz z wzrostem prędkości rośnie spalanie. Należy jednak wziąć pod uwagę, że w przypadku tira optymalna praca silnika jest zdefiniowana na prędkość około 85km/h, co jest jednym z czynników najniższej wartości spalania przy tej prędkości na wykresie. Dodatkowo warto wziąć pod uwagę, że transport odbywa się głównie poprzez drogi szybkiego ruchu, czyli każda średnia prędkość z odcinka trasy poniżej wartości w granicach 80-85km/h jest zapewne oznaką napotkanych utrudnień drogowych, które negatywnie wpływają na poziom spalania.

Ostatni z wykresów w pierwszej kolumnie reprezentuje **związek spalania od wilgotności powietrza**. W tym przypadku różnica nie jest aż tak zauważalna jak w poprzednich, jednak nadal między wartościami skrajnymi jesteśmy w stanie zaobserwować różnicę na poziomie 1,5l/100km. Biorąc to pod uwagę, zadecydowałem aby użyć tą daną w modelu.

Pierwszy z wykresów w drugiej kolumnie określa **relację pomiędzy spalaniem, a długością trasy**. Uważałem, że krótsze trasy mogą generować wyższe spalania, z uwagi na to, iż mają wyższe prawdopodobieństwo odbywania się po drogach krajowych, a nie ekspresowych, gdzie jak wcześniej zauważyliśmy spalanie wzrasta. Niestety teza ta okazała się niepoprawna i nie istnieje taka zależność. W związku z tym długość trasy nie będzie brana pod uwagę w modelu.

Drugi wykres w drugiej kolumnie natomiast wykazuje **zależność spalania od przebiegu pojazdu**. Powodem tego związku był fakt, że wraz z wzrostem przebiegu samochód ciężarowy ulega zużyciu wraz z eksploatacją. W związku z tym powinien zwiększać się opór toczenia, a w rezultacie powinno to spowodować zwiększenie spalania. W praktyce jednak na wykresie punktowym nie było możliwości dopasowania lini regresji pierwszego stopnia. Dopasowaniu uległa dopiero linia regresji trzeciego stopnia i dużymi wahaniami, z powodu czego związek ten jest nieprzydatny do mojego modelu.

Następny z wykresów prezentuje **związek spalania od pory roku**. W jego przypadku od razu możemy zauważyć, iż między wiosną a jesienią spalanie waha się na poziomie około jednego litra. Jest więc to wartościowa zależność, którą postanowiłem zawrzeć w modelu.

Ostatnie dwa wykresy to zależności jakie dodałem podczas analizy danych. Uznałem że **wpływ na spalanie może mieć pora dnia lub nocy oraz godziny szczytu.** W ten sposób powstały dwa wykresy, które ukazują wspomniane zależności. Niestety jednak okazało się, że różnice są na tyle znikome, iż żaden z tych związków nie może być zawarty w modelu.

7 Dodawanie nowych zależności

7.1 Dodanie zależności na etapie tworzenia pliku z tabelą

Jako pierwszy z dodatkowych zależności został określony atrybut "season"na etapie generowania finalnego pliku CSV. Określanie pory roku polegało na przetworzenie zmiennej "start_time"i dopasowaniu jej dnia i miesiąca do aktualnego sezonu. Przyczyną dodania tego atrybutu, był fakt iż nie możliwa byłaby do znalezienia żadna zależność od czasu, który nie był w żaden sposób pogrupowany. Dzięki dodaniu rekordu pora roku następnie łatwo było zaobserwować związki pomiędzy częścią roku a spalaniem, o czym wcześniej wspomniałem w wstępnej analizie danych.

7.2 Ustanowienie nowych związków na etapie analizy danych w Jupyter Notebook

Pomimo wstępnego pogrupowania danych czasowych postanowiłem zostawić rekordy z informacją o początku i końcu trasy w finalnej tabeli. W konsekwencji miałem możliwość dodania nowych grup czasowych na etapie analizy danych. Pierwszym z pomysłów była grupa dzieląca odcinki tras na te odbywające się między godzinami 8 i 20 oraz 20 a 8. Uważałem, iż jest to trafny podział informacji, ponieważ doszedłem do wniosku że nocą spalanie powinno być mniejsze.

Kolejną zależnością grupującą czasowe właściwości odbycia trasy jest rozdzielenie czasu jazdy na ten w godzinach szczytu oraz resztę dnia. Trafnym wydaje się zwiększenie spalania w czasie największego ruchu drogowego. Z tego powodu stworzyłem kolumnę "rush_hour".

8 Zbieranie danych treningowych i testowych do modelu

8.1 Ostateczny dobór danych do modelu

Po przeanalizowaniu wykresów doszedłem do wniosku, że do modelu użyję jedynie tych danych, które cechowały się jasno zauważalną relacją w stosunku do spalania. W ten sposób wśród nich znalazły się:

- temperatura,
- wilgotność powietrza,
- prędkość wiatru,
- średnia prędkość,
- pora roku.

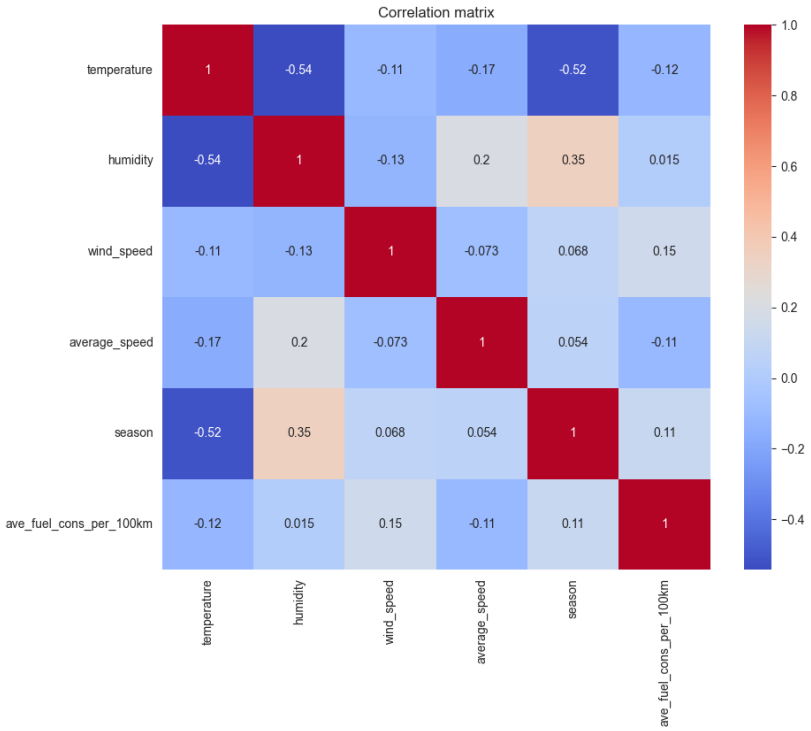
Reszta z związków miała zbyt mały lub nieliniowy wspływ na wielkość spalania, dlatego też zostały pominięte.

8.2 Podział danych na treningowe i testowe

Do podziału danych użyłem funkcji train_test_split z modułu sklearn. Podział poczyniłem w stosunku 80:20 na dane treningowe i testowe. W celu powtarzalności wyników użyłem ziarna losowości o wartości 100.

8.3 Wstępny test korelacji pomiędzy zależnościami

Sprawdzając korelację pomiędzy danymi do predykcji, a spalaniem zauważyłem, iż w we wszystkich przypadkach jest ona niestety bardzo niska. Poniżej zamieszczam tabelę z wynikami.



Rysunek 2: Macierz korelacji

Zależność spalania w żadnym przypadku nie jest wyższa niż 0.2, co oznacza bardzo słabą korelację pomiędzy nimi. W celu poprawy współczynnika spróbuję zastosować skalowanie danych.

Niestety skalowanie danych nie poprawiło korelacji. Możliwe jednak, że przyczyni się do zwiększenia jakości przewidywania modeli.

9 Tworzenie modeli

9.1 Wybór modeli

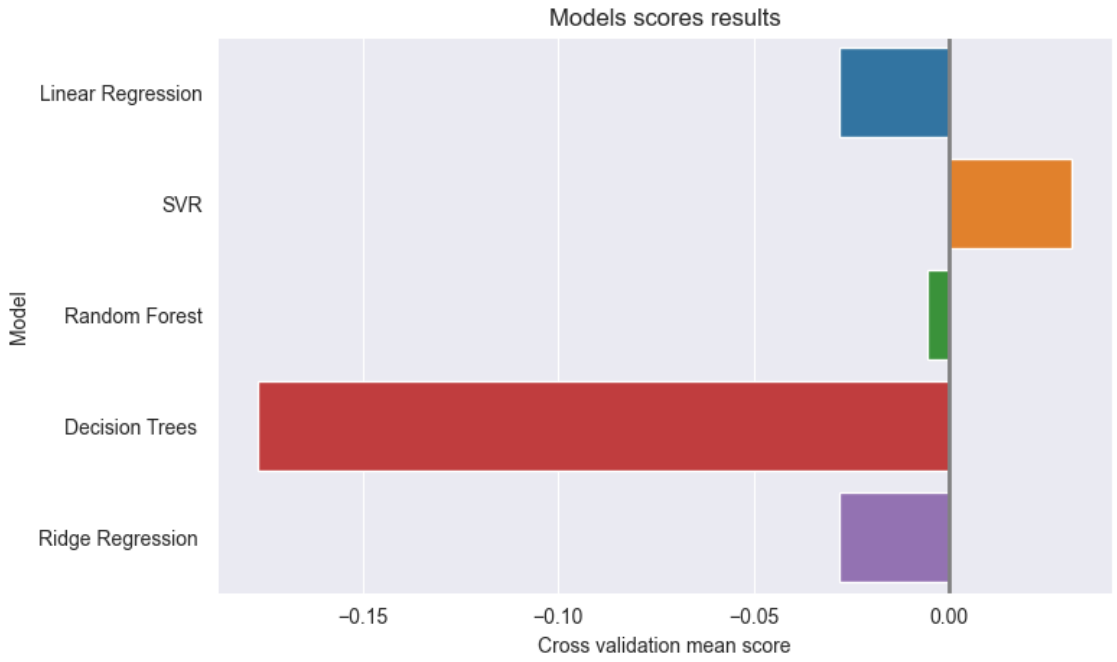
Poniżej przedstawiam modele, które wybrałem do rozwiązania.

1. Linear Regression - standardowy model regresji do problemów liniowych, jednak z powodu na niską korelację w moich danych może okazać się on nieoptymalny.
2. Support Vector Regression - model posiada zdolność do wychwytywania zależności nieliniowych, z którymi nie radzi sobie regresja, powinien dobrze sprawdzić sie w moim przypadku.
3. Random Forest Regressor - model bazujący na złożonych drzewach decyzyjnych, również posiada zdolność wychwytywania nieliniowych zależności.
4. Decision Tree Regress - model ten uczy się lokalnych regresji liniowych aproksymujących krzywą sinusoidalną, przez co wydaje odpowiedni do mojego problemu.
5. Ridge - postanowiłem sprawdzić dodatkowo ten model, jako bardziej rozbudowany względem regresji liniowej.

9.2 Przedstawienie wyników i ich omówienie

Do oceny modelów użyłem dwóch wskaźników: Root Mean Squared Error (wskaźnik pierwiastków błędów średnio-kwadratowych) oraz średnią z Cross Validation Score (walidacja krzyżowa) odpowiadający za ukazanie dopasowania modelu do danych. Poniżej w tabelach przedstawiam wyniki.

Model	Cross validation mean score
Linear Regression	-0.028220
SVR	0.031209
Random Forest	-0.002109
Decision Trees	-0.119624
Ridge Regression	-0.028194

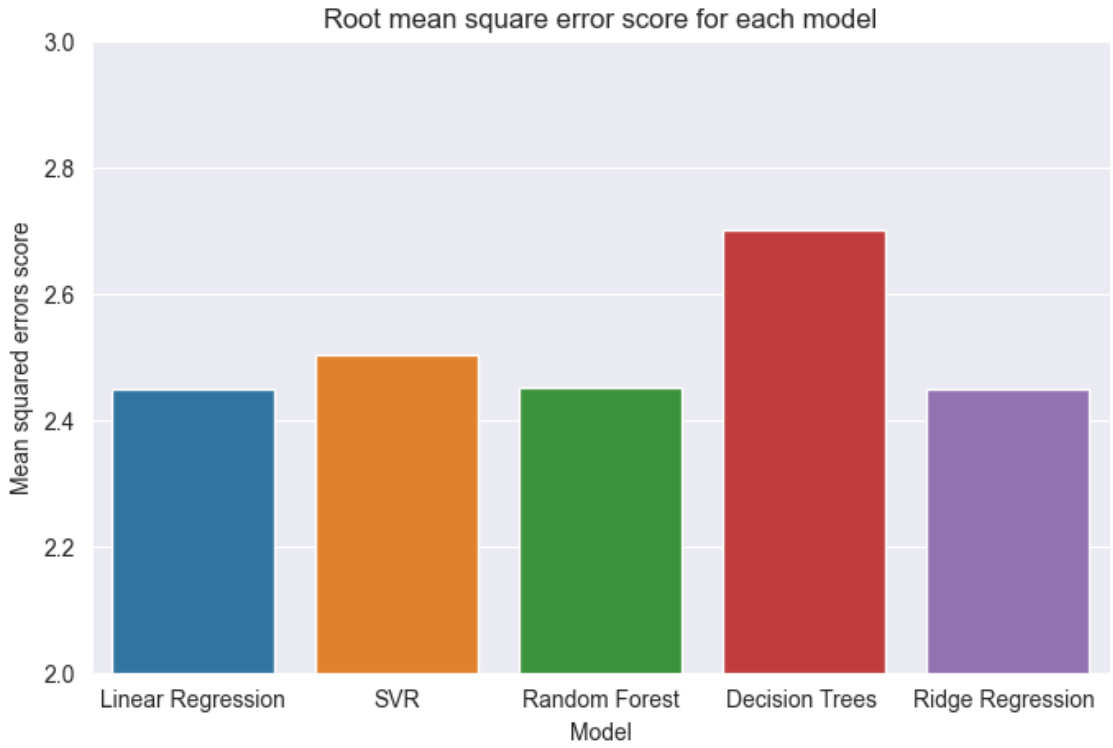


Rysunek 3: Wykres wyników średniej walidacji krzyżowej dla modeli

Ujemny wynik dla regresji liniowej oznacza, że model nie jest w stanie dopasować się do danych. Podobna sytuacja jest w pozostałych modelach z wyjątkiem SVR. Pomimo dodatniego wyniku jest on jednak bardzo słaby. W wyniku czego nie możemy zakwalifikować żadnego z modeli do określenia mianem radzących sobie z przewidywaniem wyników.

Druga z tabel reprezentuje średnią z pierwiastków błędów kwadratowych w przypadku każdego z modeli.

Model	Root mean squared errors score
Linear Regression	2.447668
SVR	2.503928
Random Forest	2.452444
Decision Trees	2.700180
Ridge Regression	2.447625

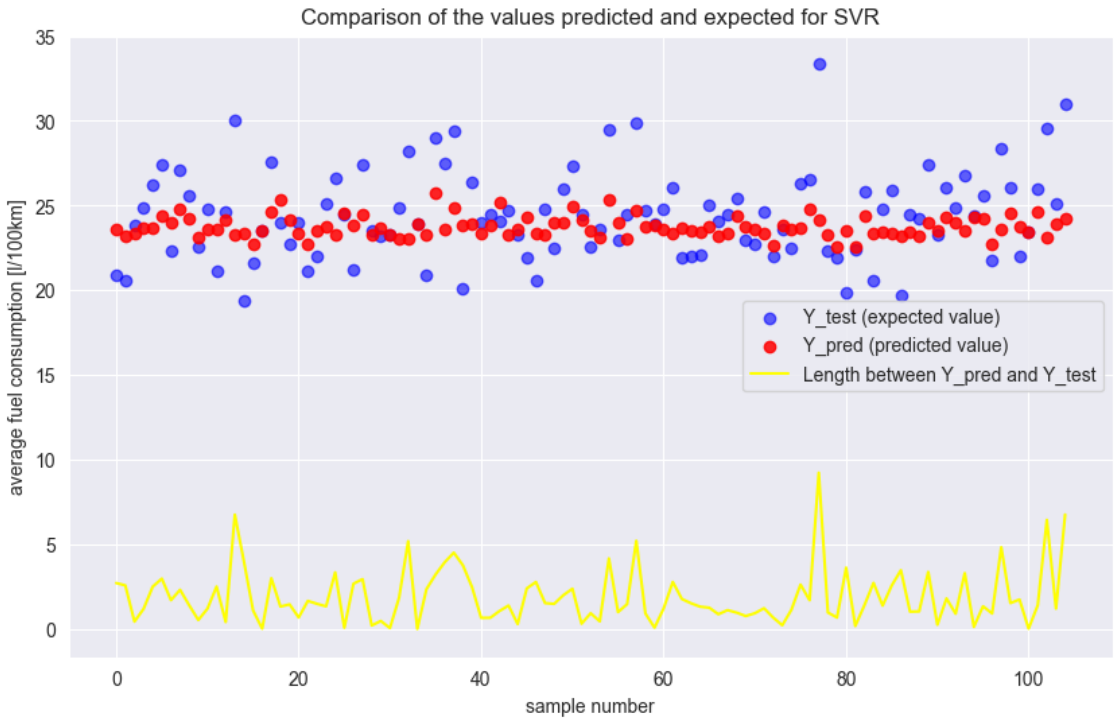


Rysunek 4: Wykres wyników pierwiastków błędów średnio-kwadratowych dla modeli

Błąd ten niestety jest stosunkowo duży i równy około 2.5 w każdym z modeli z małymi różnicami. Biorąc pod uwagę wcześniejszą analizę wykresów zależności warunków pogodowych oraz średniej prędkości wartości te są porównywalne do odchyłów spowodowanymi wspomnianymi zależnościami. W praktyce oznacza to, że zgodnie z zawartymi wykresami oraz przeprowadzoną na ich podstawie wstępną analizą różnica w spalaniu spowodowana warunkami pogodowymi wahała również się w granicach 2,5 - 3l. W rezultacie różnice te są na tym samym poziomie, co daje nam jasno do zrozumienia, że modele nie radzą sobie z odnajdywaniem różnicy w spalaniu w stosunku do pogody, albowiem margines ich błędu jest równy amplitudzie w spalaniu jaką zaobserwowaliśmy w zależności od wybranych zależności. Jednocześnie pragnę podkreślić, iż przedstawione wyniki są przed skalowaniem danych.

9.3 Przedstawienie różnicy pomiędzy wartościami oczekiwanymi, a otrzymanymi z modelów

W celu dogłębnego sprawdzenia, czy modele faktycznie słabo radzą sobie z danymi, postanowiłem stworzyć wykresy dla każdego z nich. Postaram się zawrzeć dane znajdowane przez model oraz dane oczekiwane i zobrazować odległość pomiędzy nimi. To powinno pozwolić na rozwianie wątpliwości, co do działania algorytmów. Biorąc pod uwagę jednak, że wykresy te są bardzo zbliżone, przedstawię jako przykład wykres dla SVR.

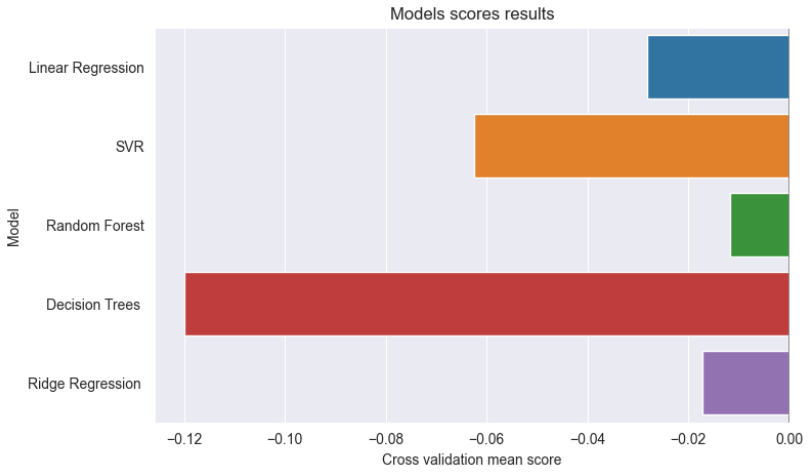


Rysunek 5: Wykres wartości predykowanych przez model w porównaniu do wartości oczekiwanych

Z wykresu możemy wyczytać słabe dopasowanie danych predykowanych do spodziewanych wartości. Oscylują one głównie wokół 25l/100km, podczas gdy wartości oczekiwane zajmują dużo większą część wykresu. Dla lepszego wychwycenia szczegółów możemy przeanalizować wykres odległości. Pokazuje on odchylenie wartości przewidzianych przez model w stosunku do testowych. W znaczącej części odchylenie to sięga 5l/100km, co jasno wskazuje na słabe dopasowanie danych przez model.

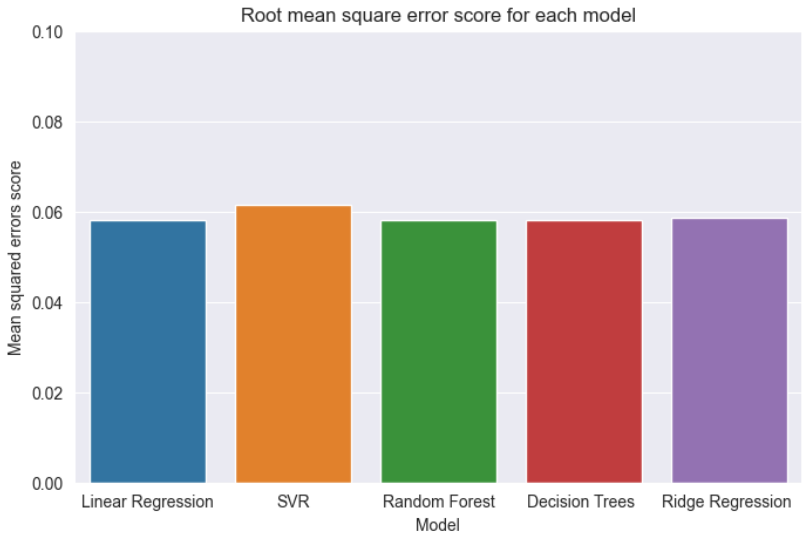
9.4 Przedstawienie wyników po skalowaniu

Niestety jak można było się spodziewać, skalowanie nie przyniosło wystarczających rezultatów. Oznacza to, że modele nadal nie są w stanie przewidywać w sposób akceptowalny.



Rysunek 6: Wykres wyników walidacji krzyżowej dla modeli po skalowaniu danych

Jak możemy zauważyć na zamieszczonym wykresie współczynnik walidacji krzyżowej nie uległ znacznemu polepszeniu dla żadnego z modeli. Dla SVR skalowanie przyniosło nawet odwrotny skutek do przewidywanego i pogorszyło jego działanie. W innych przypadkach rezultaty są podobne do poprzednich. Ujemne wartości jasno sugerują, że nie są w stanie dopasować się dodanych oraz wyciągnąć z nich wniosków.



Rysunek 7: Wykres wyników dla modeli

Na pierwszy rzut oka inaczej sprawy się mają dla współczynnika pierwiastka błędu średnio-kwadratowego, którego wartości znacząco zmalały. Mogłoby się wydawać, iż wyniki są znacząco lepsze, jednak warto wziąć pod uwagę, że MinMaxScaler z założenia przeskalowuje wyniki do wartości z zakresu [0, 1]. W tym przypadku oznacza to, że spalanie które wcześniej wynosiło liczbę rzędu 25 teraz jest w wcześniej wspomnianym zakresie. W efekcie błąd również zmalał, ponieważ względna odległość od wartości dużo mniejszych zmalała. Podsumowując, zmniejszanie wyników dla współczynnika pierwiastka błędu średnio-kwadratowego nie oznacza w tym przypadku lepszego działania modeli.

10 Wnioski

Na podstawie zebranych danych nie udało się stworzyć modelu, który byłby w stanie przewidywać wartość spalania. Za główną przyczynę takiego stanu rzeczy uważam niski współczynnik korelacji (punkt 8.3) pomiędzy zależnościami zmiennych od średniego spalania. W praktyce niska korelacja może być skutkiem innych współczynników które w bardziej znaczący sposób są skorelowane z spalaniem w samochodzie ciężarowym. Jednym z nich jest tonaż, który mógł ulec zmianą w różnych trasach. Pomimo, iż dany ciągnik zgodnie z zapewnieniami właściciela firmy transportowej przewoził głównie towar o wadze 8 ton, to niekiedy mogły występować odstępstwa, które powodowały niedeterministyczny charakter pomiarów. Dodatkową zależnością wpływającą na "zagłuszenie"danych i zmniejszenie korelacji mogła mieć również wpływ górzystość terenu. W rezultacie oznacza to, że dane pomiarowe spalania, które były prowadzone dla podobnych danych pogodowych stawały się różne na skutek wspomnianych zależności towarzyszących. Finalnie spowodowało to obniżenie korelacji dla danych, które były brane pod uwagę w modelach.

Inną przyczyną słabej przewidywalności modeli może być również zbyt mała ilość danych testowych. Niedostatecznie duży rozmiar tabeli odcinków tras nie pozwala algorytmom uczenia maszynowego na optymalne działanie. W tym przypadku dla zbioru z około 3 lat możemy wydedukować, że powtarzalność danych pogodowych jest co najmniej trój krotna, co okazuje się niewystarczające.