1 Zadanie

W naszej analizie, przyjrzymy sie zbiorowi zakwaterowań hotelowych, a naszym zadaniem bedzie przewidzieć, czy pojedyncza wizyta zostanie odwołana. Naszymi modelami uczenia maszynowego beda LightGBM oraz regresja logistyczna. Dokładność modelu LightGBM na zbiorze treningowym oraz testowym wynosza odpowiednio 89% oraz 88%.

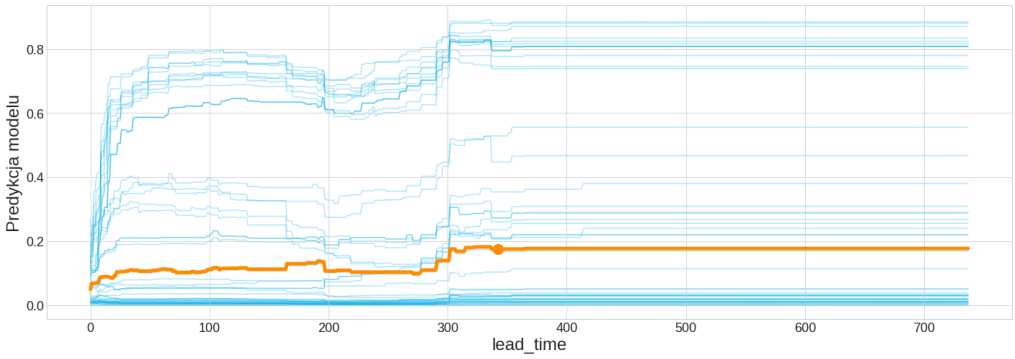
2 Zadanie

Zdaniem naszego modelu z prawdopodobieństwem 18% nasz klient odwoła rezerwacje i jest to poprawna decyzja, gdyż nasz gość tej rezerwacji nie odwołał.

3 Zadanie

Spójrzmy teraz na wykresy metody Ceteris Paribus dla kilku najciekawszych zmiennych objaśniajacych. Nasza obserwacja bedzie podkreślona kolorem pomarańczowym. Dodatkowo prawdziwa wartość zmiennej bedzie oznaczona punktem, jeżeli takiego punktu nie ma to oznacza to, że nasza obserwacja ma brakujaca wartość dla szukanej zmiennej. Pozostałe 49 losowych obserwacji dadza nam wiecej informacji o działaniu modelu z konkretna zmienna i beda oznaczone kolorem błekitnym.

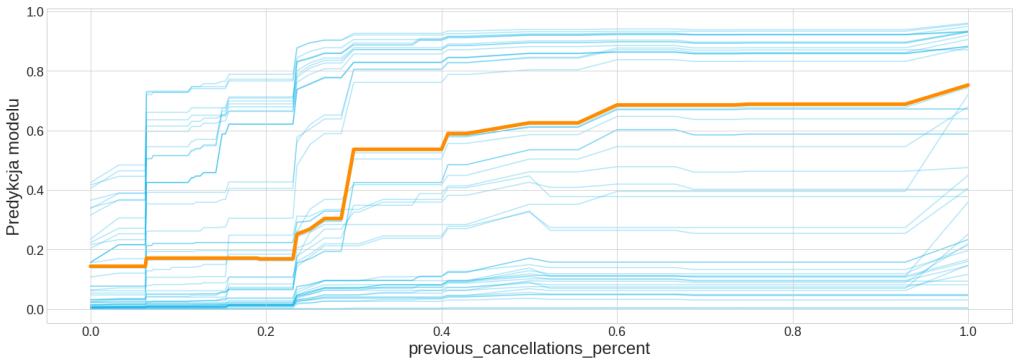
Pierwszym wykresem jest CP dla zmiennej oznaczajacej różnice miedzy data rezerwacji, a przyjazdu. Idac od prawej strony widzimy, że osoby zamawiajace zakwaterowanie z rocznym wyprzedzeniem cześciej odwołuja zamówienia. Ponadto, najwiekszy spadek odwołań jest umiejscowiony blisko miesiac przed przybyciem. Ten efekt może być zwiazany z całkowitym zwrotem kosztów jeżeli odwołamy rezerwacje odpowiednio wcześnie.



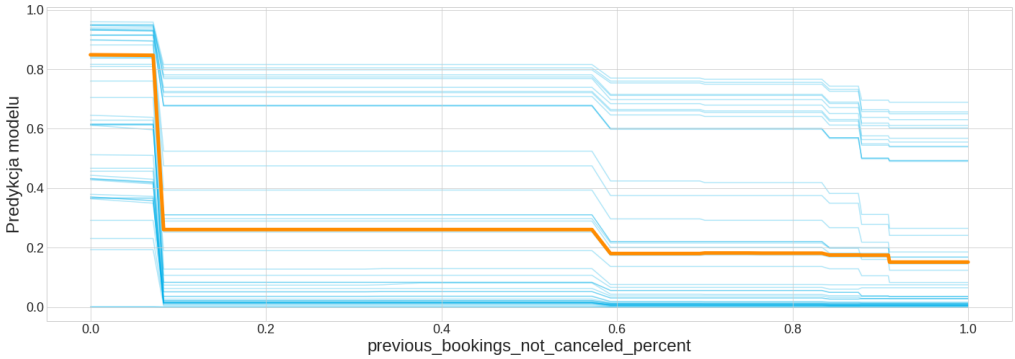
Ceteris paribus dla odległości miedzy dniem zarezerwowania, a dniem przyjazdu.

Kolejne dwie bliźniacze zmienne to procent poprzednich wizyt, które zostały

odwołane, badź nie. W obydwu przypadkach widzimy mocny trend miedzy procentem odwołanych rezerwacji, a szansa odwołania wizyty.



Ceteris paribus dla procentu poprzednich, odwołanych rezerwacji.



Ceteris paribus dla procentu poprzednich, nieodwołanych rezerwacji.

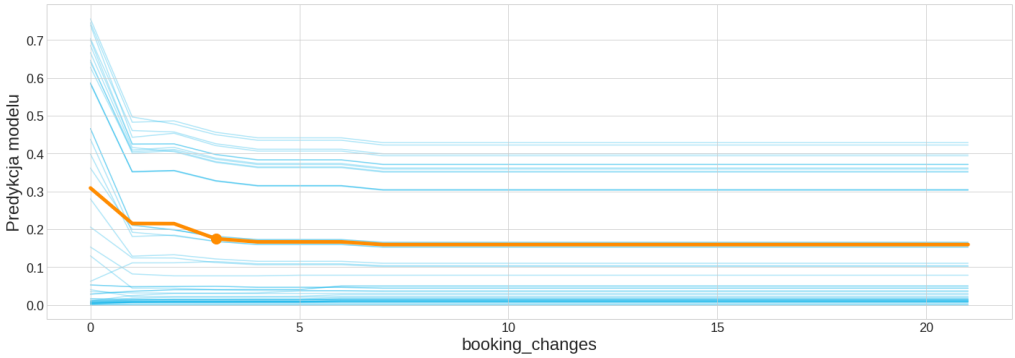
Ostatnie dwie zmienne, które rozważymy to liczba zmian w zamówieniu od poczatku jego zatwierdzenia oraz liczba miejsc parkingowych, wymaganych przez klienta. W obydwu przypadkach szansa na anulowanie pobytu drastycznie spada dla wartości wiekszych niż zero.

Dla zmiennej *booking_changes* może mieć to zwiazek z tym, że w przypadku niezadowolenia z rezerwacji mamy dwie opcje: anulowanie, albo zmiana. W przypadku, gdy prosimy o zmiany, może nam bardziej zależeć na rezerwacji.

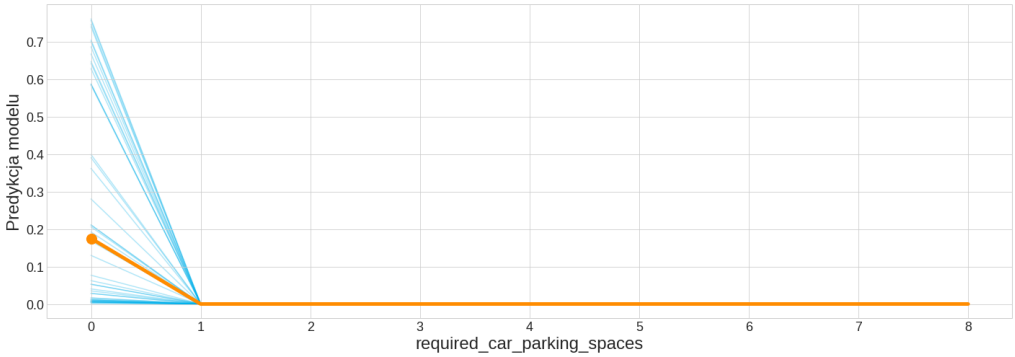
W drugim przypadku liczba miejsc parkingowych może mieć zwiazek z charakterem naszego pobytu. Jeżeli jest to pobyt w wiekszej grupie osób, to liczba miejsc parkingowych powinna wzrosnać (na przykład dla autokaru). Ponadto, rezerwacje dla wiekszej liczby osób może być bardziej przemyślana i podjeta przy konsultacji wiekszej liczby osób.

W obydwu powyższych przypadkach, dodanie zmiennych oznaczajacych czy

wyżej wymienione predykaty sa równe zero, czy nie może być przydatne przy kolejnej iteracji budowania modelu.



Ceteris paribus dla liczby zmian w zakwaterowaniu.



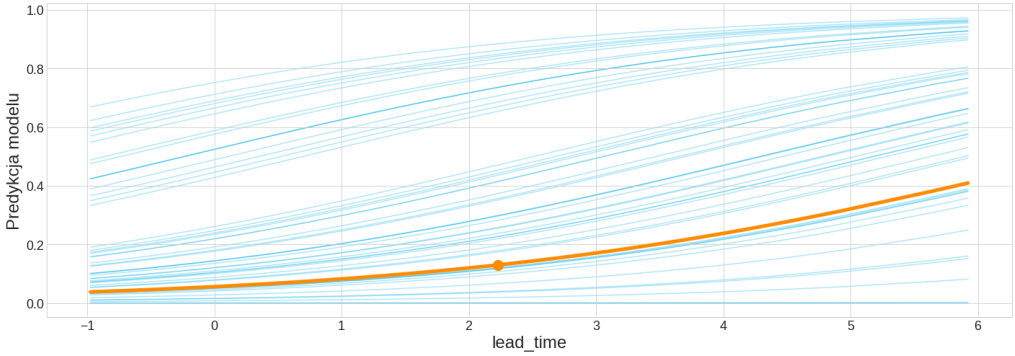
Ceteris paribus dla wymaganej liczby miejsc parkingowych.

4 Zadanie

Zmieniajac zmienna *is_repeated_guest*, zazwyczaj otrzymamy wieksze prawdopodobieństwo przy zmianie z wartości zero na jeden, gdzie predykcja modelu może sie różnić nawet o 0.4. Jest jednak przypadek, gdzie zmiana tej wartości z jedynki na zero zwieksza prawdopodobieństwo anulowania o 0.2. Powodem tego może być, niezbalansowana liczba powtórnych gości wynoszaca zaledwie 3%. Dodajac fakt, iż nasz model miał mniejszy wynik na zbiorze testowym, niż na treningowym, może oznaczać poczatek przeuczania sie modelu.

5 Zadanie

To co możemy zauważyć w modelu logistycznym to gładkie wykresy CP, co oczywiście jest zwiazane z tym, że decyzja naszego modelu jest funkcja ciagła, różniczkowalna. Najwieksze różnice można zauważyć dla zmiennej *lead_time*. W przypadku użycia metod drzewiastych widoczny jest punkt przegiecia funkcji, gdzie z wykresu wklesłego zmienia sie on na wypukły, podczas gdy w nowym modelu tego nie ma. Nie mniej jednak, sens wyjaśnienia tej zmiennej pozostaje prawdziwy (Osoby, które wcześniej zamawiaja, cześciej anuluja rezerwacje). Jednak w przypadku prostego modelu liniowego, lokalne zmiany nie zostały wyłapane. Dodanie zmiennych wielomianowych od tej zmiennej może poprawić powyższy model logistyczny.



Ceteris paribus modelu regresji logistycznej dla odległości miedzy dniem zarezerwowania, a dniem przyjazdu.