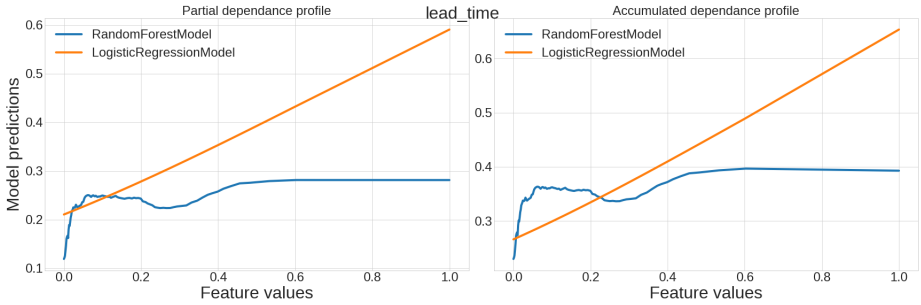
XAI - Praca domowa #6

Miłosz Kacper Michta

May 14, 2020

W naszej analizie, przyjrzymy sie zbiorowi zakwaterowań hotelowych, a naszym zadaniem bedzie przewidzieć, czy pojedyncza wizyta zostanie odwołana. Naszymi modelami uczenia maszynowego beda las losowy oraz regresja logistyczna.

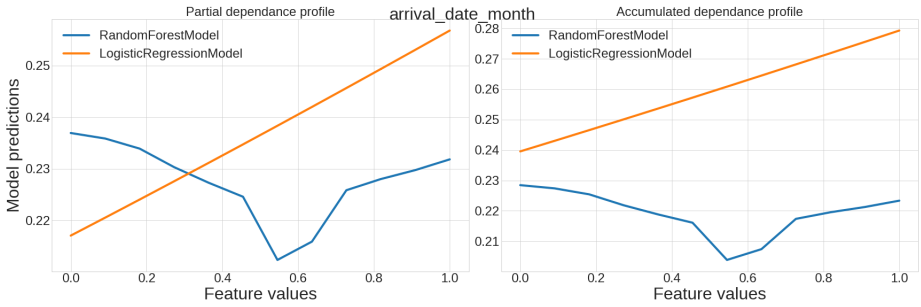
Spójrzmy teraz na profile czastkowe i kumulatywne, naszych modeli pod katem kilku cech. Zaczniemy od zmiennej *lead_time* oznaczajacej różnice w dniach miedzy data zakwaterowania, a rezerwacji.



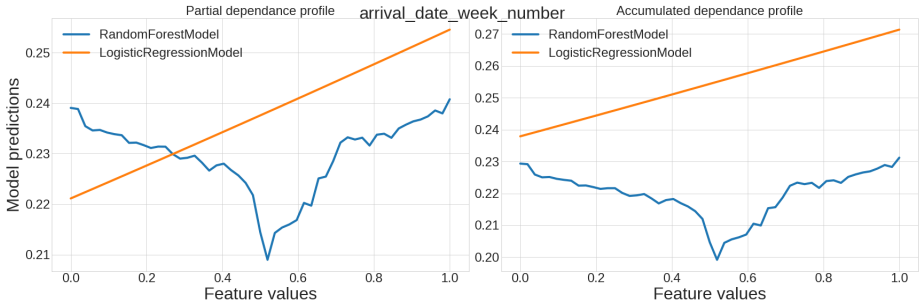
Profile dla odległości miedzy dniem zarezerwowania, a dniem przyjazdu.

To co możemy zaobserwować patrzac na ten wykres, to przewaga modelu drzewiastego, który jest w stanie lepiej sie dopasować do danych z nieliniowym trendem. Kolejna różnica jest to, że dla opcji kumulatywnej, predykcje obu modeli przyjmuja wieksze wartości, co mówi nam o tym, że modele zawieraja interakcje, które osłabiaja wpływ powyższej cechy samej w sobie.

Kolejne dwie bliźniacze zmienne to czasy przybycia do hotelu, odpowiednio w miesiacach i tygodniach.



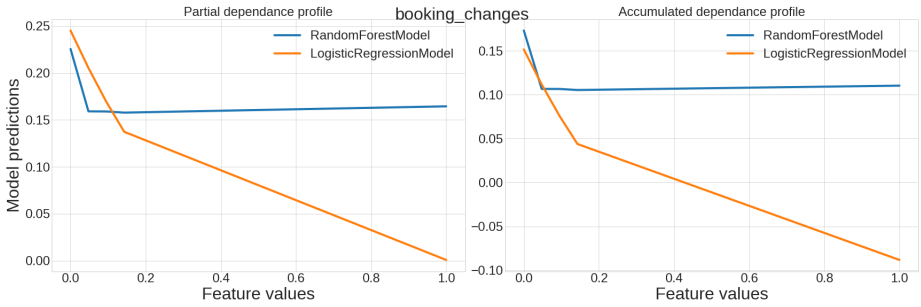
Profile dla miesiaca zakwaterowania.



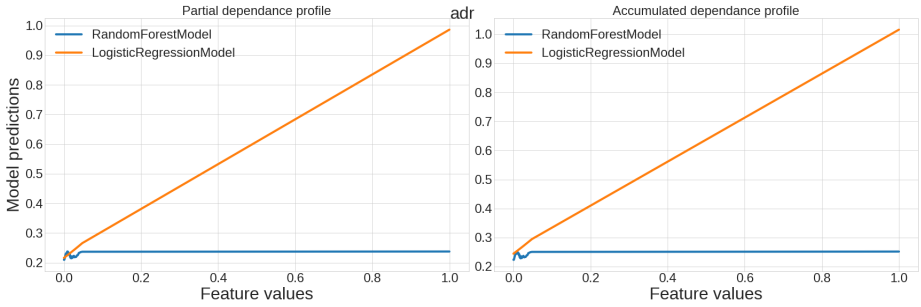
Profile dla tygodnia zakwaterowania.

Podobnie jak w poprzednim przykładzie, model drzewiasty ma przewage nad modelem logistycznym w postaci wykonania podziału na danych niemonotonicznych/nieliniowych. Uwzgledniajac wakacyjna pore roku, las losowy uznał, że w tym okresie prawdopodobieństwo anulowania rezerwacji jest mniejsze z uwagi na zajetość miejsc w wiekszości wakacyjnych kurortów.

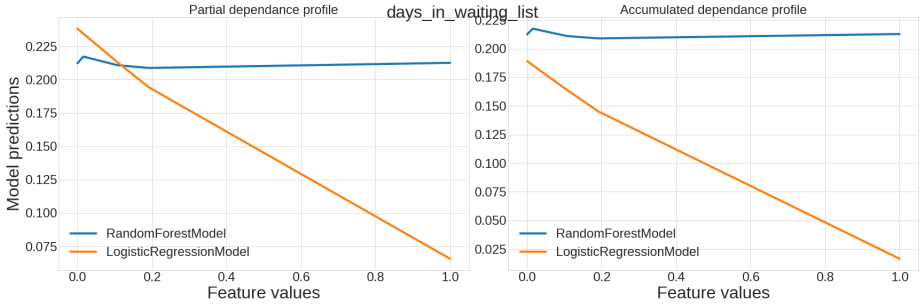
W powyższych przykładach, mogliśmy zaobserwować duża skuteczność modelu drzewiastego, jednak poniżej zaprezentujemy przykłady zmiennych, które przechylaja szale w strone modelu regresji logistycznej.



Profile dla liczby zmian w zamówieniu.



Profile dla średniej ceny za pokój w wybranym hotelu.



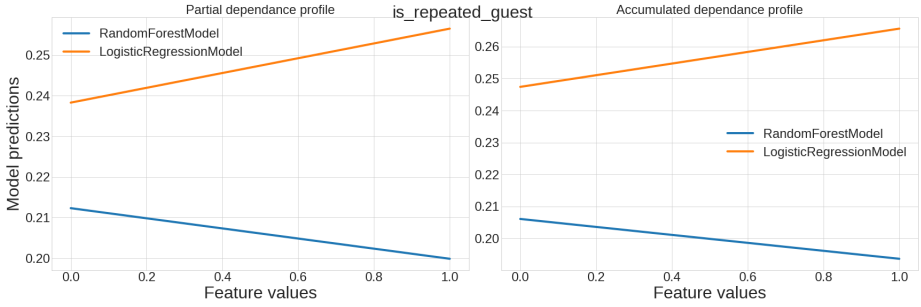
Profile dla długości oczekiwania na akceptacje, liczona w dniach.

Pierwsza rzecza rzucajaca sie w twarz, to brak reakcji lasu losowego na zmiany wartości w powyższych zmiennych, w przeciwieństwie do regresji logistycznej.

Wszystkie trendy w pokazanych, powyżej grafach maja sensowna interpretacje. Jeżeli składamy zmiany w rezerwacji, to zapewne zależy nam na danym zakwaterowaniu. Jesteśmy bardziej skłonni zrezygnować z drogiego wywczasu,

niż z atrakcyjnej oferty. Im dłużej czekamy na zwolnienie miejsca w hotelu, tym wieksze jest nasze zaangażowanie.

Jednak dlaczego model drzewiasty nie jest zdolny nauczenia sie tych kilku bardzo prostych reguł? Wyjaśnień może być wiele, jedno z nich to fakt, że modelom drzewiastym zdecydowanie trudniej jest aproksymować funkcje liniowe, niż modelom liniowym. Kolejnym bardzo znaczacym powodem (moim zdaniem) moga być pierwsze trzy przykłady, które obrazuja jak las losowy wybrał cechy zwiazane z data jako bardziej istotne, niż zmienne ciagłe opisujace rezerwacje. Spójrzmy jeszcze na różnice przy zmienna opisujaca czy ktoś jest powracajacym gościem.



Profile dla cechy *is_repeated_guest*.

To co możemy zaobserwować to, wzajemnie przeczace sobie profile dla zmiennej *is_repeated_guest*. Model drzewiasty poprawnie znajduje trend w odróżnieniu od regresji logistycznej. Ponadto, negatywny trend modelu liniowego ma zdecydowanie mniejszy wpływ na model patrzac na profil czastkowy, a kumulatywny jeszcze bardziej pogłebia te różnice. Powodem tego może być niezbalansowanie przykładów dla danej zmiennej (gości powracajacych jest około 3%), przez co maksymalizacja entropii przy podziale może mieć wieksze znaczenie, niż logistyczna funkcja straty.

To co można wykorzystać z powyższej analizy, to wzmocnienie modelu logistycznego poprzez dodanie zmiennych wielomianowych, badź innego kodowania zmiennych opisujacych date (chociażby, przez kodowanie cykliczne na okregu jednostkowym).