Porównanie najlepszych modeli z 3 użytych metod

	Random Forest	SVM	KNN
Dokładność	0.852	**0.866**	**0.866**
Czułość	**0.845**	0.831	0.831
Specyficzność	0.859	**0.897**	**0.897**
Avg Dokładność CV	**0.908**	0.877	0.874
Avg Czułość CV	**0.867**	0.849	0.866

Modele SVM i KNN osiągnęły takie same wyniki na zbiorze testowym oraz bardzo podobne przeciętne wyniki podczas testowania ich wydajności za pomocą Cross Validation, z tą różnicą, że model KNN poradził sobie lepiej pod kątem przeciętnej czułości, a model SVM odrobinę lepiej pod kątem przeciętnej dokładności.

Modele SVM i KNN osiągnęły też dokładność większą o 1.4 pkt % od modelu Random Forest, natomiast las losowy wykazał się wyższą czułością o 1.4 pkt %. Przeciętne lepsze wyniki modelu testowane za pomocą CV osiągnął również model Random Forest. Modele SVM i KNN mimo próby maksymalizacji czułości przy zachowaniu jak najwyższej dokładności i tak przewyższają dokładność i czułość wynikami specyficzności.

Pomimo wyższej ostatecznej dokładności, modele odległościowe sprawdziły się w tutaj nieco gorzej niż las losowy pod kątem jednoczesnej maksymalizacji czułości i dokładności, oraz szczególnie wyników testowanych za pomocą CV.

Bazując na zebranych informacjach, przy próbach dalszej optymalizacji jakości modelu można skupić się na użyciu innych technik decyzyjnych, np. boostujących takich jak AdaBoost, GradientBoost czy XGBoost.

Analiza interpretowalności modelu Random Forest

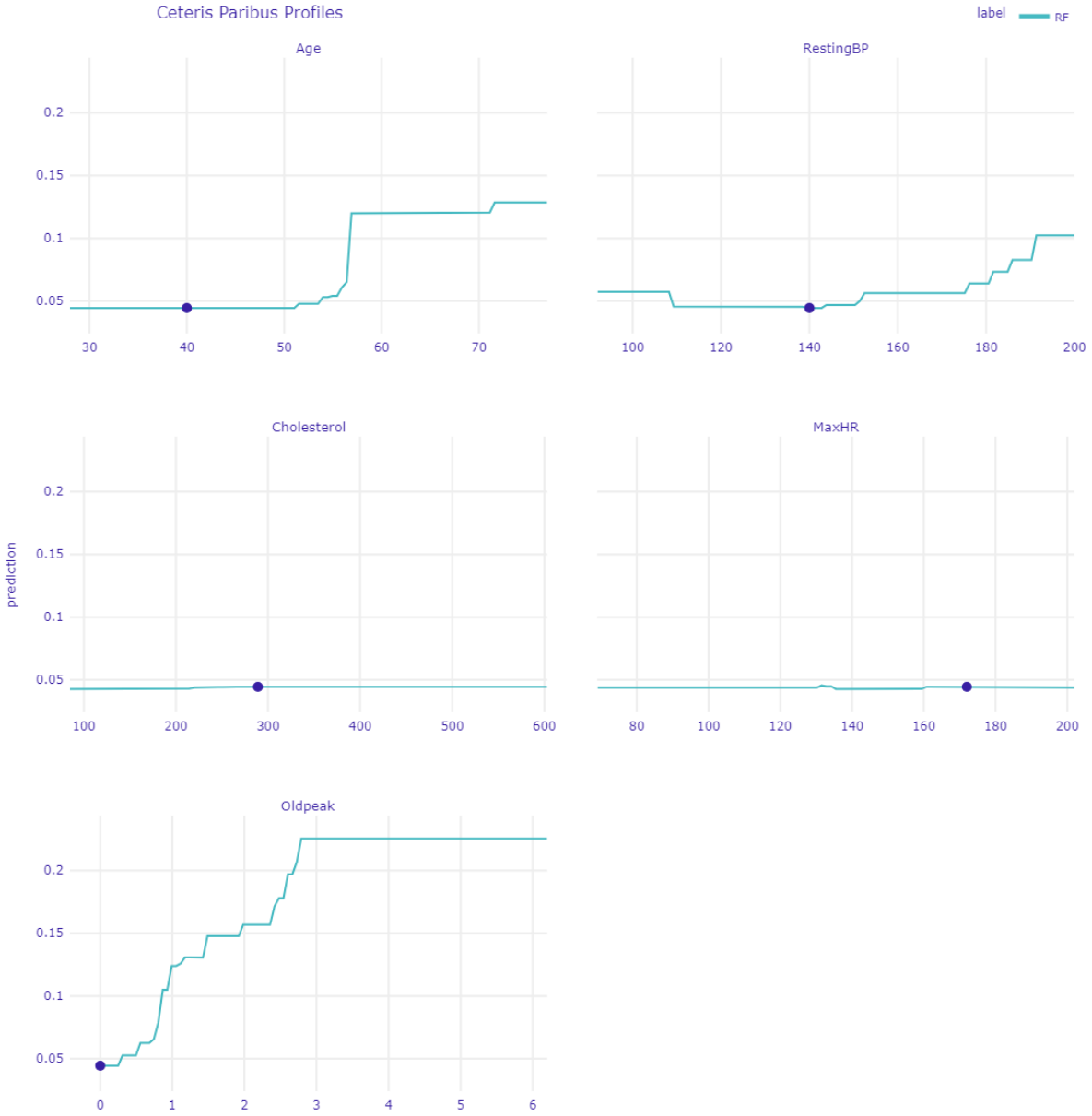
Profile ceteric-paribus

Zbadano wpływ zmiennych objaśniających na zmiany predykcji wywołane zmianami wartości tych zmiennych za pomocą profili ceteris-paribus dla pierwszej obserwacji ze zbioru danych. Oto jej statystyki:

Age	40.0
RestingBP	140.0
Cholesterol	289.0
FastingBS	0.0
MaxHR	172.0
Oldpeak	0.0
Sex_M	1.0
ChestPainType_ASY	0.0
ChestPainType_ATA	1.0
ChestPainType_NAP	0.0
ChestPainType_TA	0.0
RestingECG_LVH	0.0
RestingECG_Normal	1.0
RestingECG_ST	0.0
ExerciseAngina_Y	0.0
ST_Slope_Down	0.0
ST_Slope_Flat	0.0
ST_Slope_Up	1.0
HeartDisease	0.0

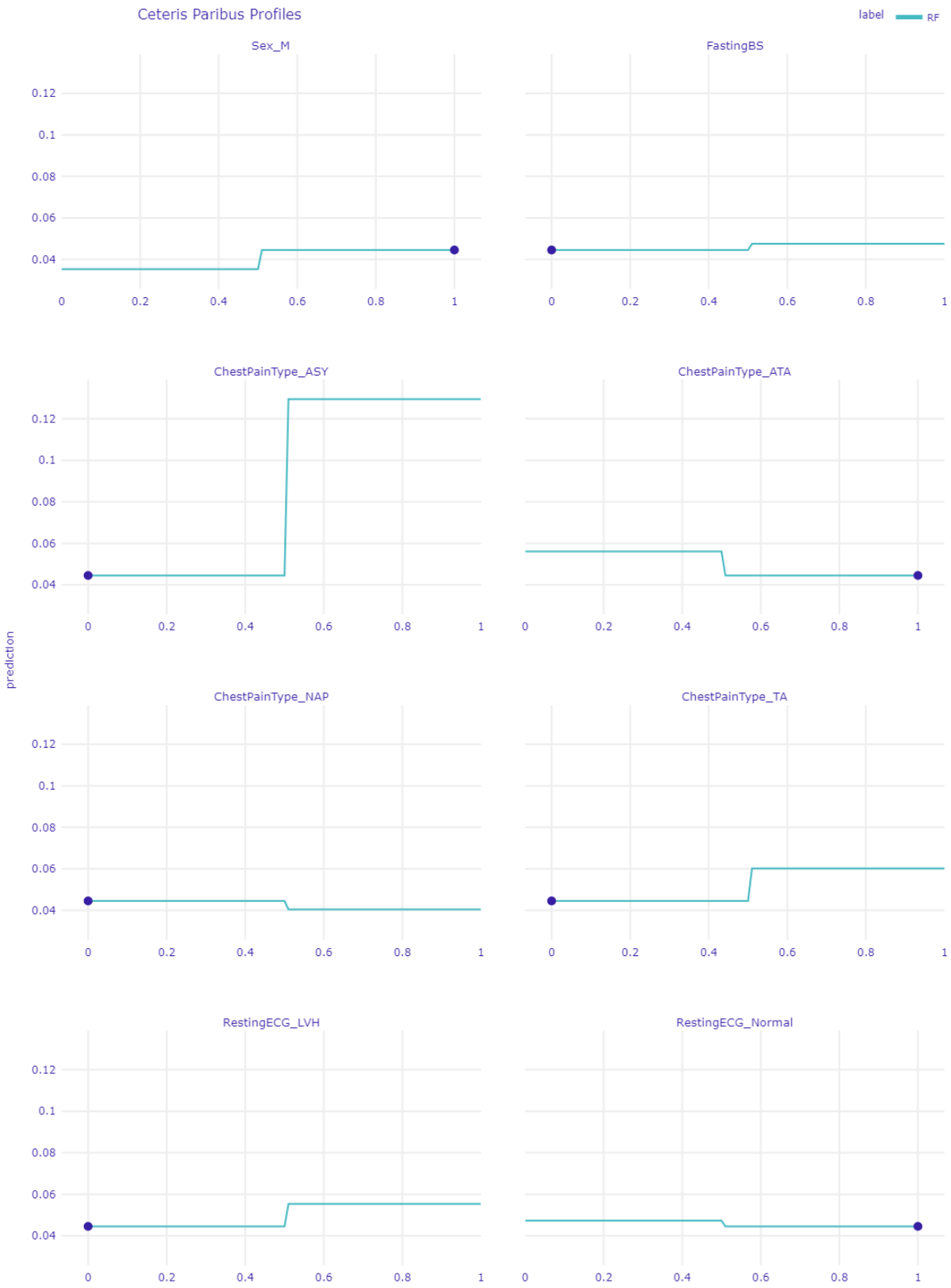
Jak widać jest to mężczyzna w wieku 40 lat ze zdrowym sercem. Nie występowały u niego objawy potencjalnie wskazujące na występowanie choroby serca. Jednym negatywnym czynnikiem może być dość wysoki cholesterol pacjenta.

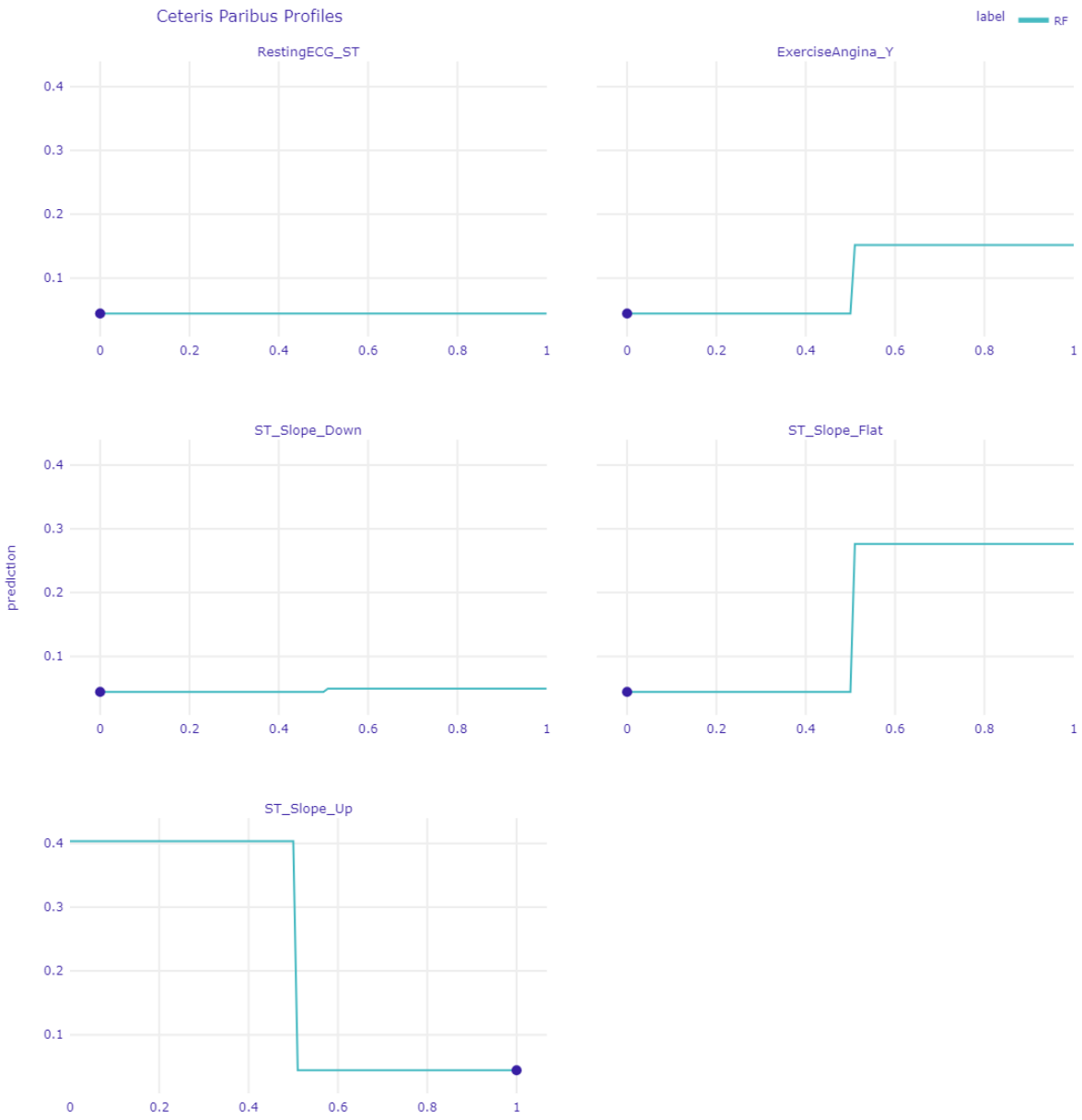
Oto jak prezentują się profile ceteris-paribus dla zmiennych numerycznych:



Można zauważyć gwałtowny wzrost prawdopodobieństwa klasyfikacji badanego pacjenta jako osoba potencjalnie chora na serce przez nasz model po osiągnięciu wieku 57 lat. Dodatkowo, skrajnie wysokie wartości spoczynkowego ciśnienia krwi mogłyby zwiększyć prawdopodobieństwo klasyfikacji jako osoba chora o jakieś 5 pkt %. Wśród zmiennych numerycznych, największy wpływ na klasyfikację badanego pacjenta ma zmienna Oldpeak, która od wartości 2.8 daje 22.6% szans na wystąpienie choroby serca. Zmiany wartości Cholesterol czy MaxHR nie miałyby takiego wpływu na klasyfikację pacjenta przez model.

Profile ceteris-paribus dla zmiennych kategorycznych wyglądały następująco:

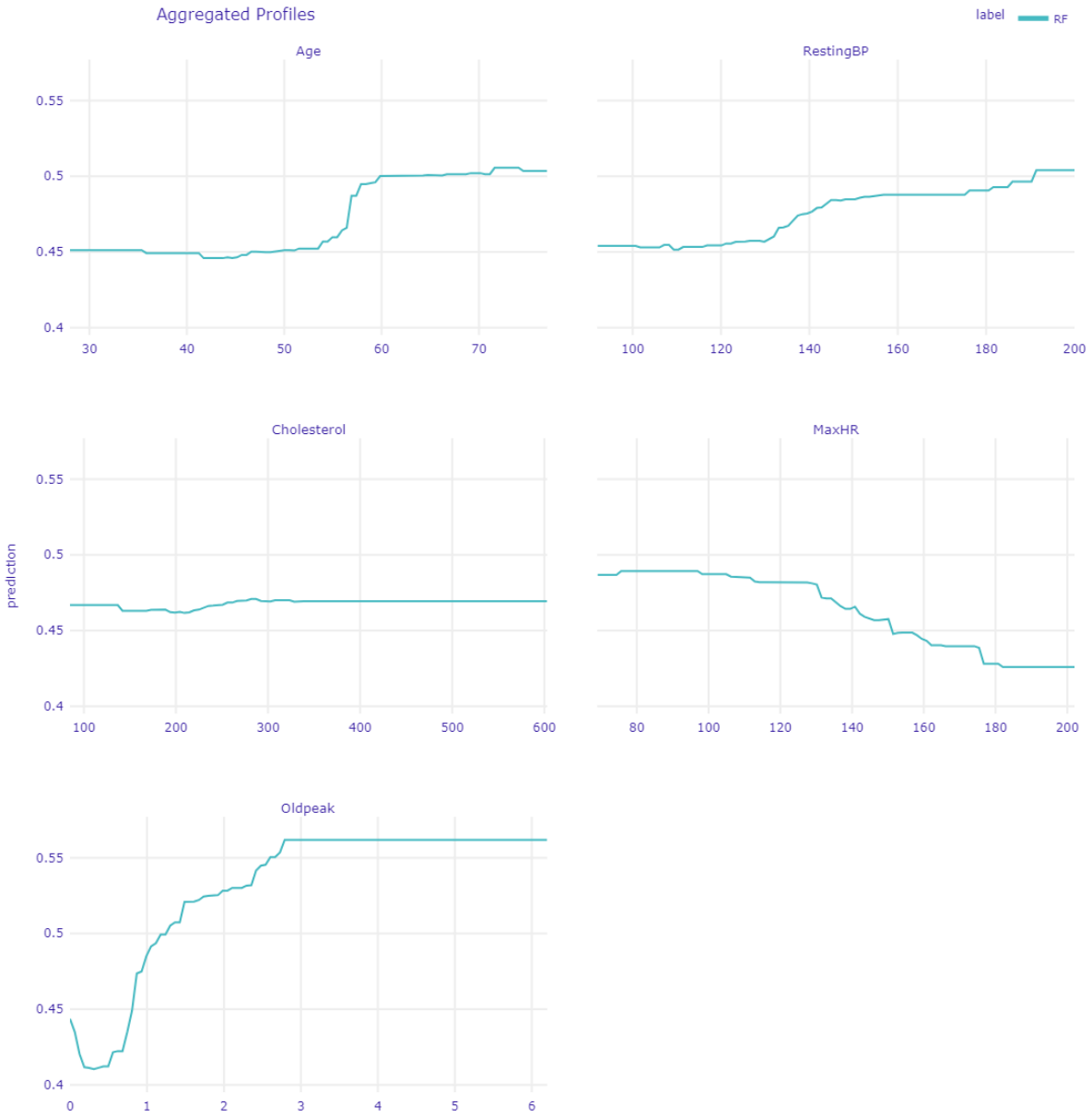




Gdyby nasz pacjent był kobietą, jego szansa na wystąpienie choroby serca wg naszego modelu zmalałaby o 0.9 pkt %. Zmienna ST_Slope_Up ma największy wpływ na klasyfikację naszego pacjenta: brak posiadania tej cechy zwiększył by prawdopodobieństwo wystąpienia choroby serca o 36 pkt % wg naszego modelu. ST_Slope_Flat było też ważną cechą, której posiadanie zwiększa prawdopodobieństwo klasyfikacji jako osoba z chorobą serca o 23.2%. Innymi ważnymi wskaźnikami było ExcerciseAngina_Y oraz ChestPainType_ASY. Pozostałe zmienne nie miały aż tak wyraźnego wpływu na wynik klasyfikacji badanego pacjenta przez nasz model.

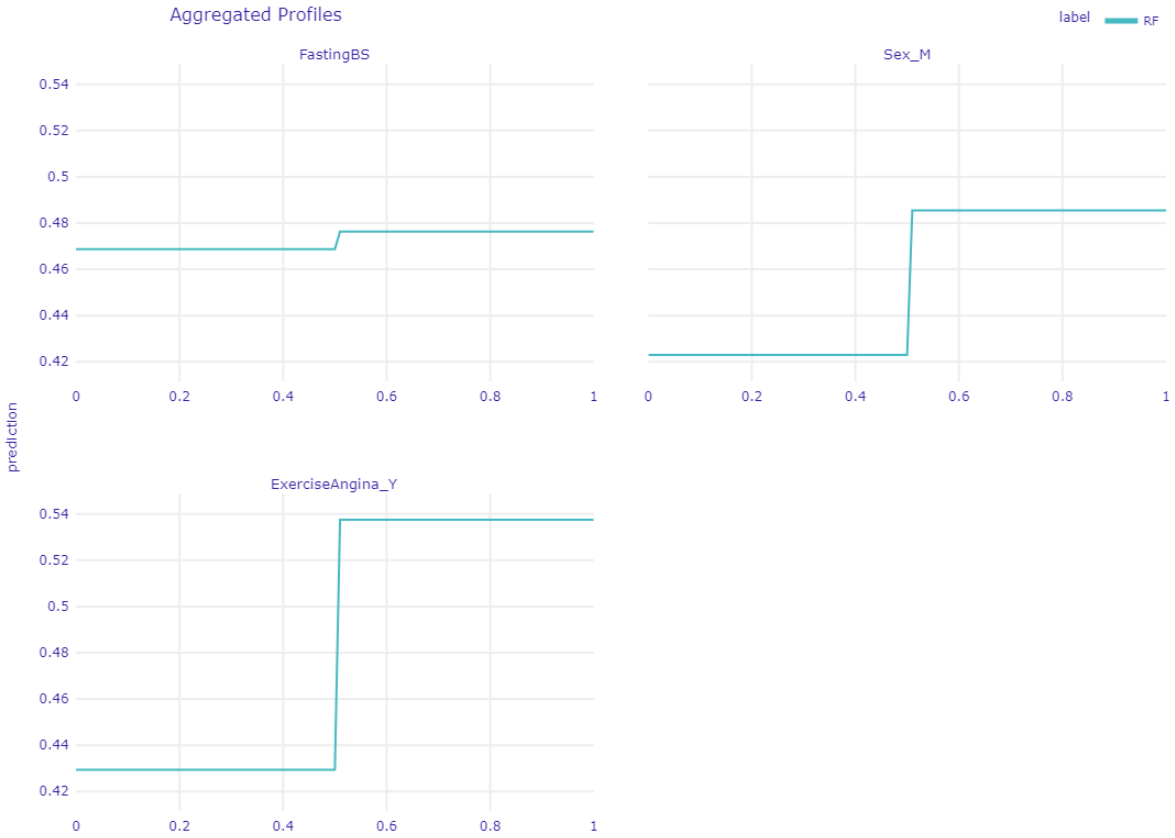
Wykresy częściowej zależności

Utworzono wykresy częściowej zależności, które informują nas o oczekiwanej wartości predykcji modelu w zależności od wartości zmiennych jako średnie z indywidualnych profili ceteris paribus badanej populacji. Wyglądają one następująco dla zmiennych numerycznych:



W porównaniu z wcześniej badaną osobą, zdecydowaną różnicą we wpływie zmiennych na klasyfikację możemy zaobserwować dla MaxHR. Jej skrajnie wysokie wartości w znaczącym stopniu obniżają prawdopodobieństwo klasyfikacji jako osoba chora na serce. Poziom cholesterolu nadal nie wpływa z znaczącym stopniu na wyniki klasyfikacji modelu. Spoczynkowe ciśnienie krwi wykazało się większym wpływem wśród całej populacji niż w przypadku badanej wcześniej osoby.

Wpływ zmiennych kategorycznych na klasyfikację modelu przedstawiony zostanie osobno dla zmiennych kategorycznych binarnych i zmiennych o kilku kategoriach.



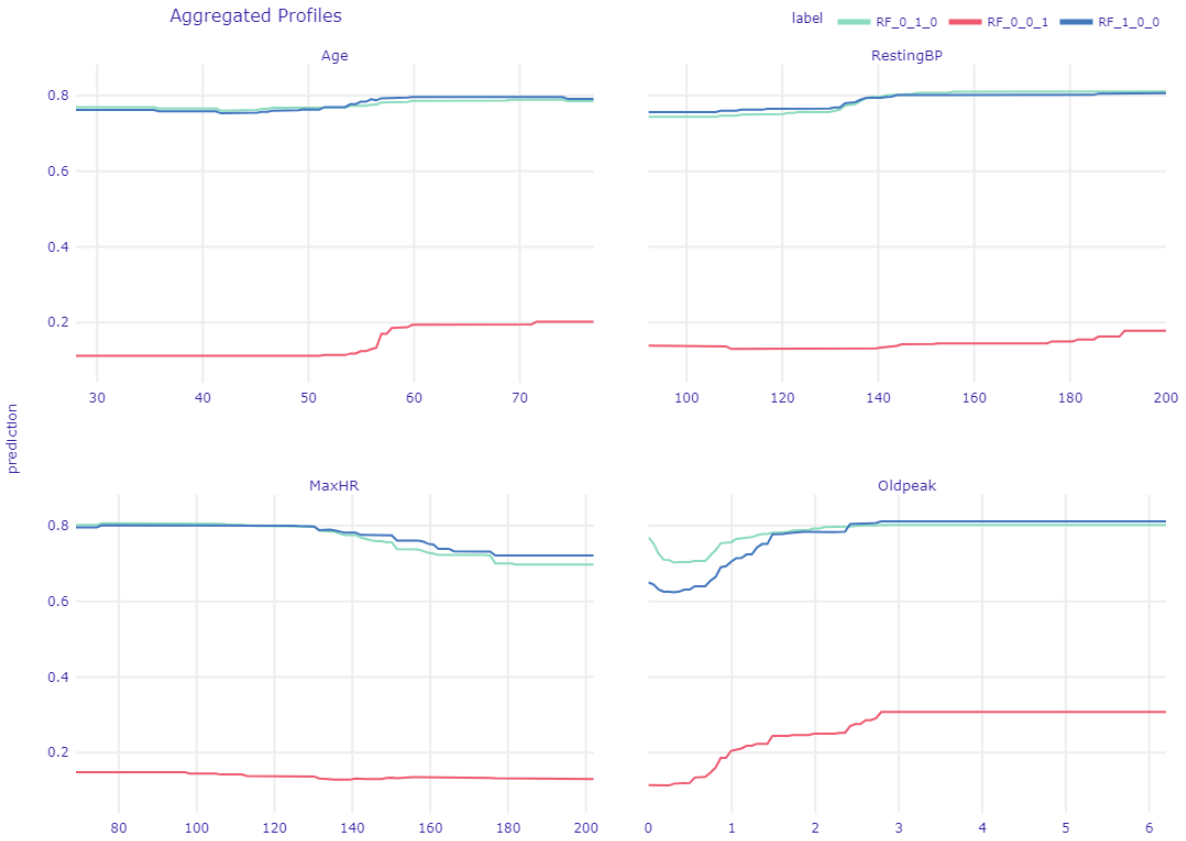
Wśród zmiennych kategorycznych binarnych, duży wpływ na klasyfikację ma płeć oraz występowanie dławicy wywołanej wysiłkiem fizycznym. Mężczyźni będą klasyfikowani jako chorzy częściej niż kobiety przez nasz model.

Wpływ zmiennych kategorycznych posiadających ponad 2 kategorie pokazany zostanie w przekroju z 4 najważniejszymi zmiennymi numerycznymi: Age, RestingBP, MaxHR oraz Oldpeak.

Oto jak wartości ST_Slope wpływają na klasyfikację przez nasz model, w kombinacji z 4 najważniejszymi zmiennymi numerycznymi.

Profile dla ST_Slope:

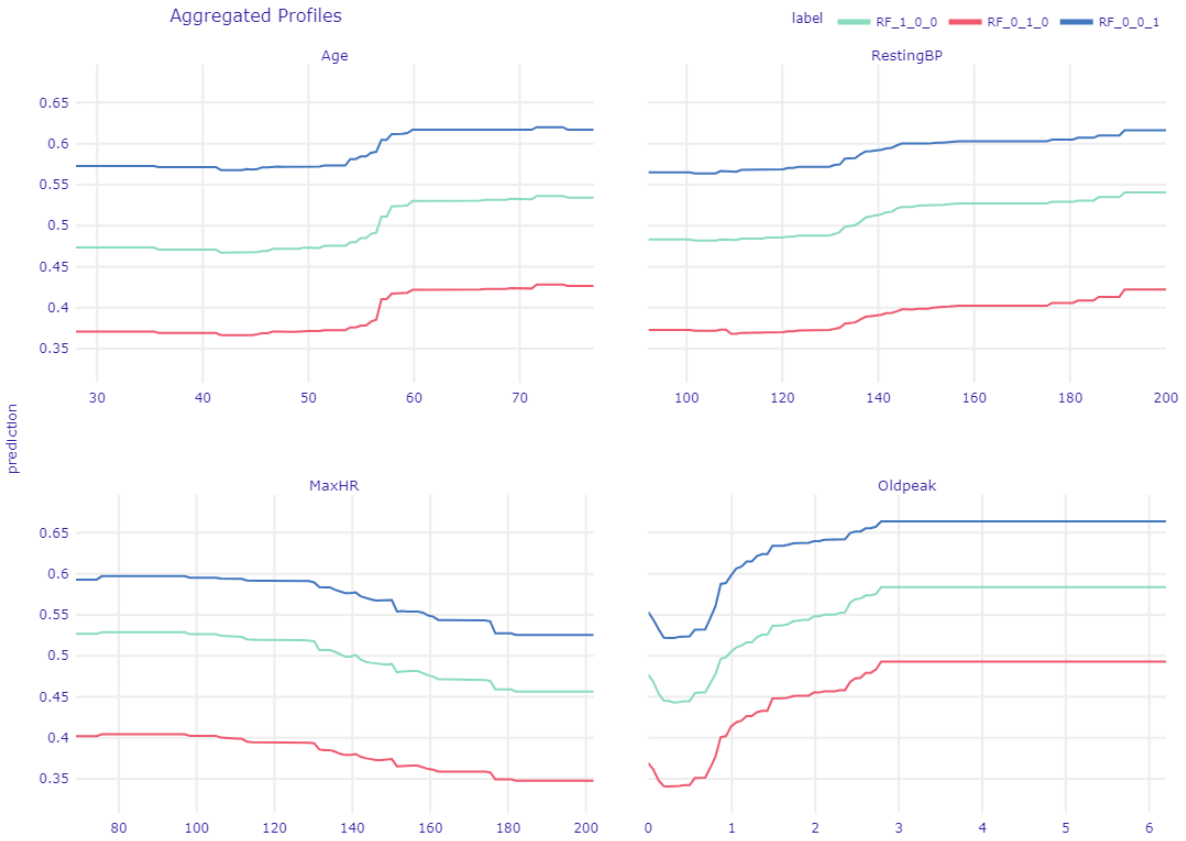
0_1_0 – Flat, 0_0_1 – Up, 1_0_0 - Down



Wartość ST_Slope Up w porównaniu do Flat i Down zmniejsza prawdopodobieństwo klasyfikacji jako osoba chora na serce. Dla wartości Age można zauważyć większy wpływ zmiennej Age na wynik klasyfikacji kiedy zmienna ST_Slope przyjmuje wartość up. Dla wartości Flat oraz Down rosnące RestingBP w większym stopniu wpływa na dodatnie prawdopodobieństwo jako osoba chora. Wysokie wskazania MaxHR dla tych wartości bardziej wpływają na klasyfikację jako osoba zdrowa niż przy ST_Slope Up, ze względu na już bardzo niskie wartości przewidywanej klasyfikacji. Dla zmiennej Oldpeak wykresy dla każdej z kategorii ST_Slope różnią się wzajemnie najbardziej. Dla osób z cechami ST_Slope Flat oraz Down zerowe i bliskie zeru wartości Oldpeak zwiększają prawdopodobieństwo klasyfikacji jako osoba chora w porównaniu do wartości Oldpeak około 0.2 do 0.7.

Profile dla RestingECG:

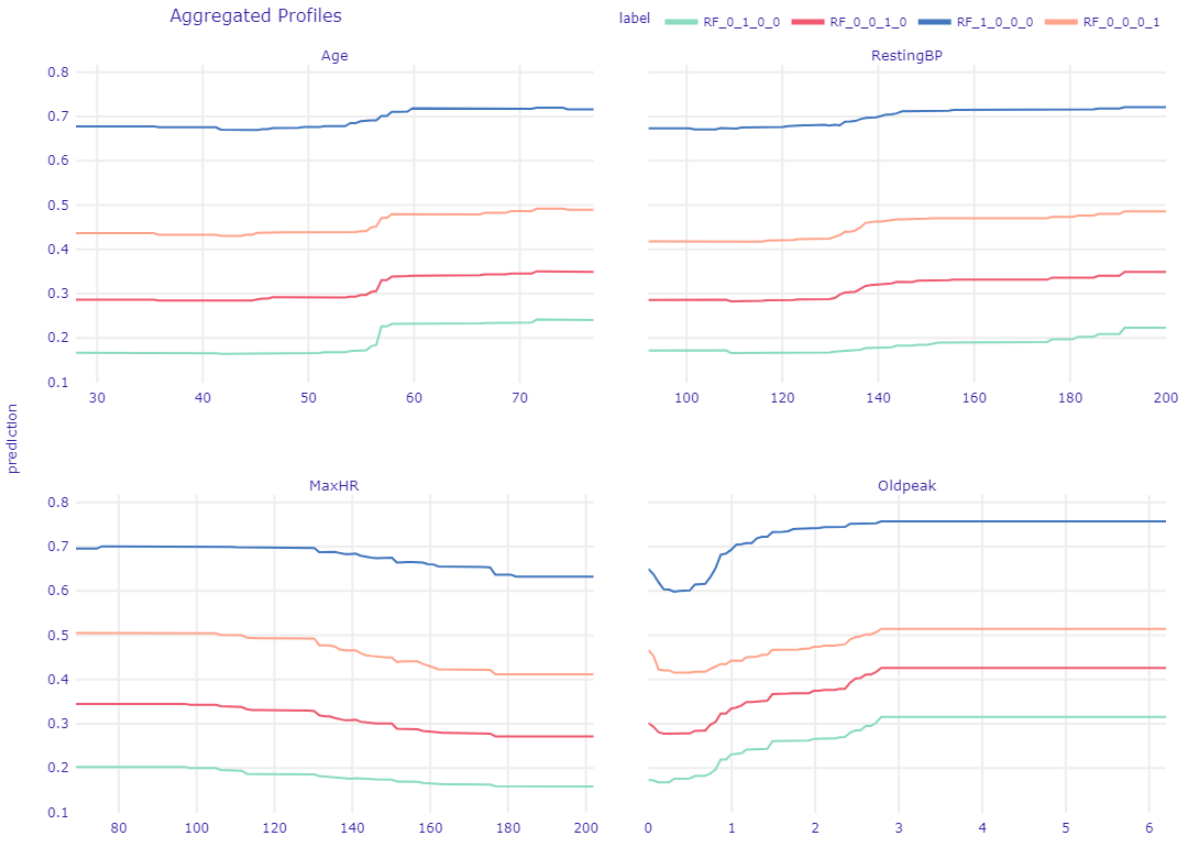
0_1_0 – Normal, 0_0_1 – ST, 1_0_0 - LVH



Przynależność do którejkolwiek kategorii RestingECG nie wpływa znacząco na kształtowanie się wykresów częściowej zależności niezależnie od zmiennej numerycznej. Osoby z normalnym odczytem ECG są najmniej narażone na wystąpienie choroby serca wg naszego modelu. Osoby z cechą LVH są bardziej narażone, natomiast najbardziej osoby z cechą ST.

Profile dla ChestPainType:

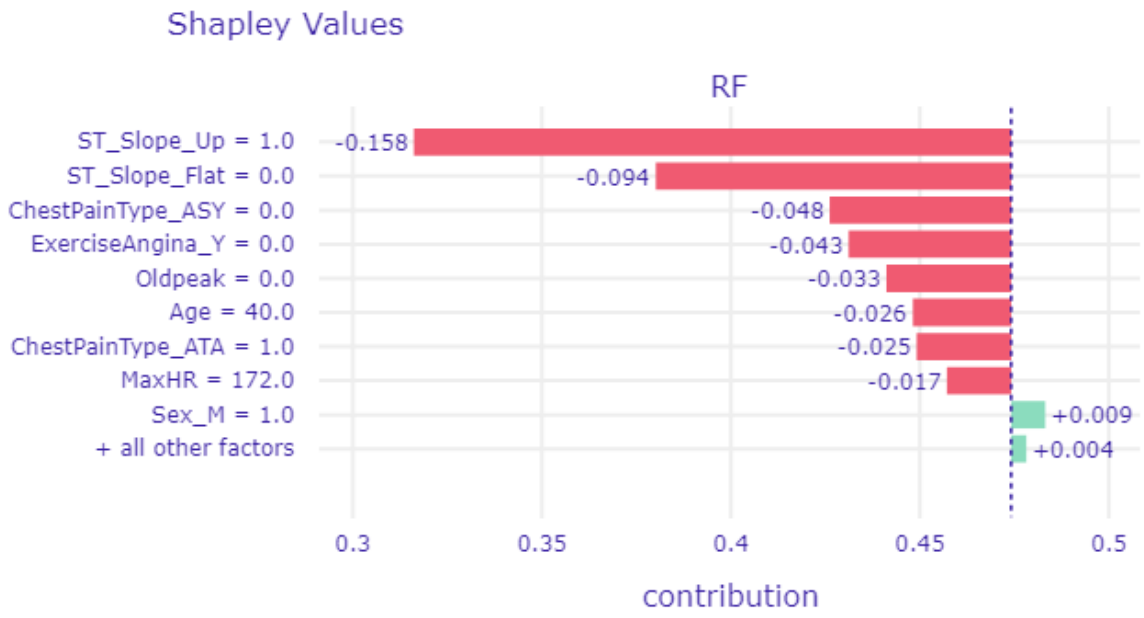
0_1_0_0 – ATA, 0_0_1_0 – NAP, 1_0_0_0 – ASY', 0_0_0_1 - TA



Dla zmiennej kategorycznej ChestPainType osoby asymptotyczne były najczęściej klasyfikowane jako chore w porównaniu z innymi grupami. Najbezpieczniejsza jest grupa osób ATA. Tak jak wcześniej, widzimy inne zakrzywianie się wykresów dla zmiennej Oldpeak w okolicach zera, gdzie dla cech bardziej narażonych na chorobę serca Oldpeak równe 0 w coraz większym stopniu wpływa na pozytywną klasyfikację jako osoba chora.

Wartości SHAP

Zbadamy teraz, jak cechy wcześniej badanej osoby person1 wpłyną na jej klasyfikację przez model.

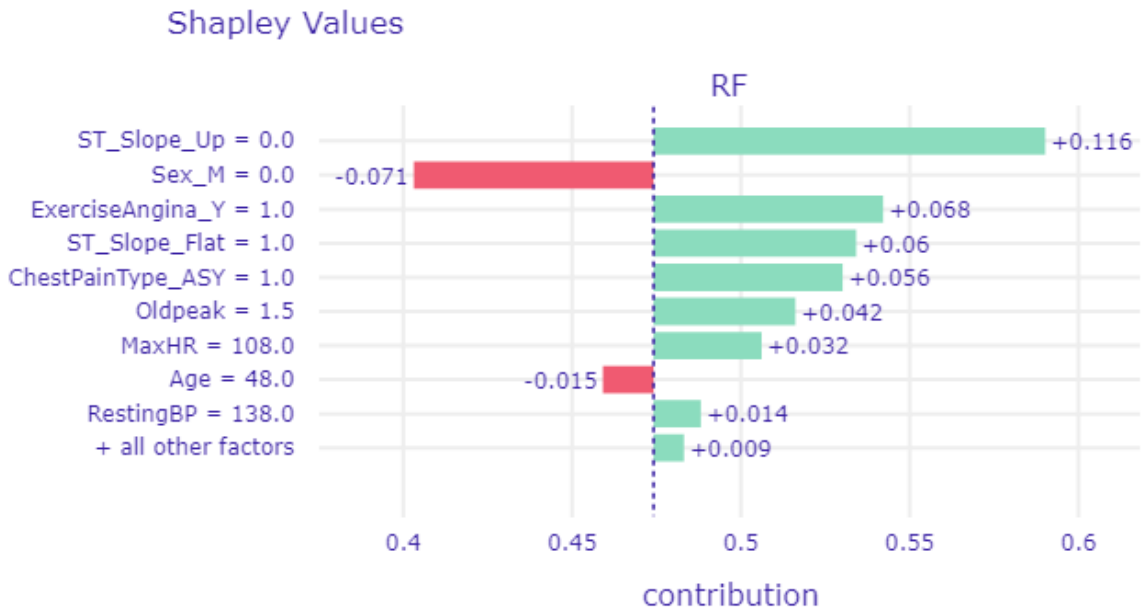


Najważniejszym czynnikiem wpływającym na negatywną klasyfikację pacjenta jest zdrowy odczyt odcinka ST, tzn. wystąpienie cechy ST_Slope Up i przez to też nie wystąpienie ST_Slope Flat. Brak asymptotycznego bólu w klatce piersiowej i nie wystąpienie dławicy oddechowej po aktywności fizycznej też w dużym stopniu kontrybuują do negatywnej klasyfikacji pacjenta. Zerowa wartość Oldpeak i niska wartość wieku, tzn. 40 lat, kiedy osoba nie jest jeszcze aż tak podatna na wystąpienie choroby serca są kolejnymi czynnikami negatywnie wpływającymi na klasyfikację.

Zobaczymy teraz, jak cechy osoby z chorobą serca o słabym profilu zdrowotnym będą wpływać na jej prawdopodobieństwo klasyfikacji. Oto statystyki porównywanej pacjentki:

Age	48.0
RestingBP	138.0
Cholesterol	214.0
FastingBS	0.0
MaxHR	108.0
Oldpeak	1.5
HeartDisease	1.0
Sex_M	0.0
ChestPainType_ASY	1.0
ChestPainType_ATA	0.0
ChestPainType_NAP	0.0
ChestPainType_TA	0.0
RestingECG_LVH	0.0
RestingECG_Normal	1.0
RestingECG_ST	0.0
ExerciseAngina_Y	1.0
ST_Slope_Down	0.0
ST_Slope_Flat	1.0
ST_Slope_Up	0.0

Badaną osobą będzie 48 letnia kobieta. Osiągnęła ona bardzo niskie maksymalne tętno, dodatkowo wartość Oldpeak nie wyniosła w jej przypadku 0, wystąpiła u niej dławica wywołana aktywnością fizyczną, St_Slope przjęło wartość flat.



Czynnikiem wpływającym najmocniej na klasyfikację jako osoba chora jest niewystąpienie cechy ST_Slope Up. Płeć pacjentki w negatywny sposób wpłynęła na prawdopodobieństwo klasyfikacji jej jako osoba chora przez model. Wystąpienie cech ExerciseAngina_Y, ChestPainType_ASY oraz ST_Slope Flat wpłynęło znacząco przewidywaną klasyfikację jako osoba chora. Oldpeak na poziomie 1.5 oraz niskie maksymalne odczytane tętno wpłynęły też na przewidywaną klasyfikację jako osoba chora.