Streszczenie

Eksploracja medycznych zbiorów danych często wymaga zastosowania analizy przeżycia, obszaru statystyki, w którym interesuje nas przewidzenie czasu do wystąpienia określonego wydarzenia, np. pojawienia się choroby, niewydolności narządowej, czy śmierci. Istotną częścią analizy przeżycia jest uwzględnianie tzw. obserwacji cenzurowanych, czyli podmiotów, u których nie zaobserwowano wystąpienia obserwowanego wydarzenia w okresie obserwacji. Występowanie obserwacji cenzurowanych wymaga zastosowania specjalnych metod modelowania statystycznego, których wyprowadzenie należy do obszaru analizy przeżycia.

W niniejszej dysertacji zgłębiony został temat metod uczenia maszynowego, których rezultatem są modele przeżycia pozwalające przewidzieć czas do wystąpienia określonego wydarzenia. Metody te są istotnym narzędziem pozwalającym na odkrycie medycznych czynników ryzyka oraz dającym lepsze rezultaty predykcyjne na heterogenicznych i cenzurowanych zbiorach danych medycznych w porównaniu z tradycyjnymi metodami modelowania statystycznego, takich jak model proporcjonalnego ryzyka Coxa. W pracy tej dodatkowo uwagę poświęcono interpretowalności modeli, która jest szczególnie istotna w stosowanej analizie medycznej, gdyż zrozumienie i transparentność działania modeli jest konieczne, aby były one wykorzystane w zastosowaniach klinicznych, kiedy mamy do czynienia z sytuacjami mogącymi wpłynąć na życie ludzkie.

Niniejsza praca ma następujący układ. Rozdział 1 przedstawia problem oraz tezy badawcze rozprawy. Rozdział 2 wprowadza zagadnienia i notacje obszaru analizy przeżycia, w tym popularny model proporcjonalnego ryzyka Coxa oraz miarę jakości predykcyjnej modeli przeżycia – C-index. Rozdział 3 przedstawia sposoby dostosowania metod uczenia maszynowego do problematyki analizy przeżycia oraz opisuje dwie popularne metody z tego obszaru – lasy losowe przeżycia (ang. *random survival forests*) i sieci neuronowe przeżycia (ang. *survival neural networks*).

W Rozdziale 4 wyprowadzona została nowatorska metoda uczenia maszynowego dla modeli przeżycia, która pozwala ona osiągnięcie wysokiej mocy predykcyjnej porównywalnej z modelami o pełnej złożoności przy zachowaniu pełnej interpretowalności modeli, tj. możliwości określenia wpływu każdego predyktora na wynik. Osiągnięto to modyfikując metodę wzmocnienia gradientowego w wariancie składnikowym (ang. *component-wise gradient boosting*) i dostosowując ją do problemów analizy przeżycia poprzez optymalizację funkcji straty w postaci błędu średniokwadratowego przeżycia. Dodatkowo wykorzystane

zostały modele bazowe (ang. *base learners*) w postaci zespołu drzew (ang. *bagged trees*) przy ograniczeniu wynikowej formy funkcyjnej modelu do sumy addytywnych funkcji nieliniowych pobierających na wejściu pojedynczy predyktor lub ich parę. Rozdział 5 zawiera empiryczną ewaluację komponentów zaproponowanej metody oraz jej porównanie na rzeczywistych zbiorach danych z innymi metodami uczenia maszynowego dostosowanymi do analizy przeżycia, pokazując wysoką moc predykcyjną zaproponowanej metody.

Rozdział 6 zawiera rezultat klasycznej medycznej analizy statystycznej przeprowadzonej na niepublicznym zbiorze danych pacjentów po przeszczepieniu wątroby. Zaproponowany model oparty jest na modelu proporcjonalnego ryzyka Coxa i wykorzystuje statyczną wartość AST, zmienność PLT oraz trend PLT i WBC. Daje on lepsze wyniki od ugruntowanego modelu MELD (ang. *Model for End-Stage Liver Disease*) oraz modeli zbudowanych tylko z wykorzystaniem statycznych wartości pomiarów biochemicznych. Wynikająca z modelu informacja, że zmienność i trend PLT mierzonego w okresie pierwszego roku po przeszczepieniu wątroby są istotnymi predyktorami długoterminowej przeżywalności pacjentów, jest wykorzystywana w praktyce przez lekarzy. Rozdział 7 zamyka rozprawię poprzez podsumowanie zawartości, przedstawienie głównych wyników pracy oraz potwierdzenie udowodnienia postawionych tez pracy.