Statistical Learning - Projekt zaliczeniowy RAPORT

Celem analizy danych przeprowadzonej w naszym projekcie jest stworzenie modelu oceniającego, jakie czynniki mają wpływ na poziom szczęścia społeczeństwa. Pierwszy tego typu raport został opublikowany w 2012 roku i z biegiem czasu zyskał coraz większe zainteresowanie. Z biegiem czasu raport zdobył coraz większe uznanie na świecie, ponieważ rządy, firmy i społeczeństwo obywatelskie coraz częściej korzystają ze wskaźników szczęścia do podejmowania swoich decyzji w sprawach np. polityki. Wiodący eksperci z różnych dziedzin - ekonomii, psychologii, analizy ankietowej, statystyki narodowej, zdrowia, polityki publicznej i innych - opisują, jak pomiary dobrostanu mogą być skutecznie wykorzystywane do oceny postępu państw. Raporty oceniają stan szczęścia na świecie dzisiaj i pokazują, jak nowoczesna nauka o szczęściu wyjaśnia różnice indywidualne i narodowe w poziomie szczęścia.

Krok 1. Zbieranie danych

Zbiór danych użyty w projekcie pochodzi ze strony kaggle i można go znaleźć pod linkiem:

https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2022.csv&fbclid=IwAR03IeQIyVRH-bVk5DbX4qAEGILujHLFdm1eOvKo3IzSA8WyJPetf3_0y-M

Zbiór zawiera Wskaźnik Szczęścia dla 153 krajów wraz z czynnikami wyjaśniającymi tę wielkość. Zgromadzone wyniki wykorzystują dane z Ogólnoświatowego Sondażu Gallupa. Punkty są oparte na odpowiedziach na pytanie dotyczące oceny życia, zadane w ankiecie. To pytanie, znane jako schody Cantrila, prosi respondentów o ocenę swojego obecnego życia w skali 0-10.

Podstawowy zbiór danych zawiera 12 kolumn:

- RANK - pozycja kraju w rankingu ogólnej oceny szczęścia. RANK wskazuje, jak dany kraj plasuje się w porównaniu do innych,

- Country - kraj, w którym przeprowadzono badania,
- Happiness.score - wskaźnik ogólnego poziomu szczęścia w danym kraju,
- Whisker.high i Whisker.low - przedziały ufności dla oceny szczęścia. Wskazują na zakres, w którym mieści się prawdziwa wartość poziomu szczęścia,
- Dystopia..1.83….residual - kraj teoretyczny, którego wartości są równe najniższym narodowym średnim dla każdej z sześciu zmiennych. Jest to pewnego rodzaju punkt odniesienia, który pomaga porównywać kraje względem tego, jak dobrze im się powodzi w porównaniu do sytuacji dystopijnej,

W tym zmienne objaśniające; czynniki, które wpływają na ogólny poziom szczęścia w kraju:

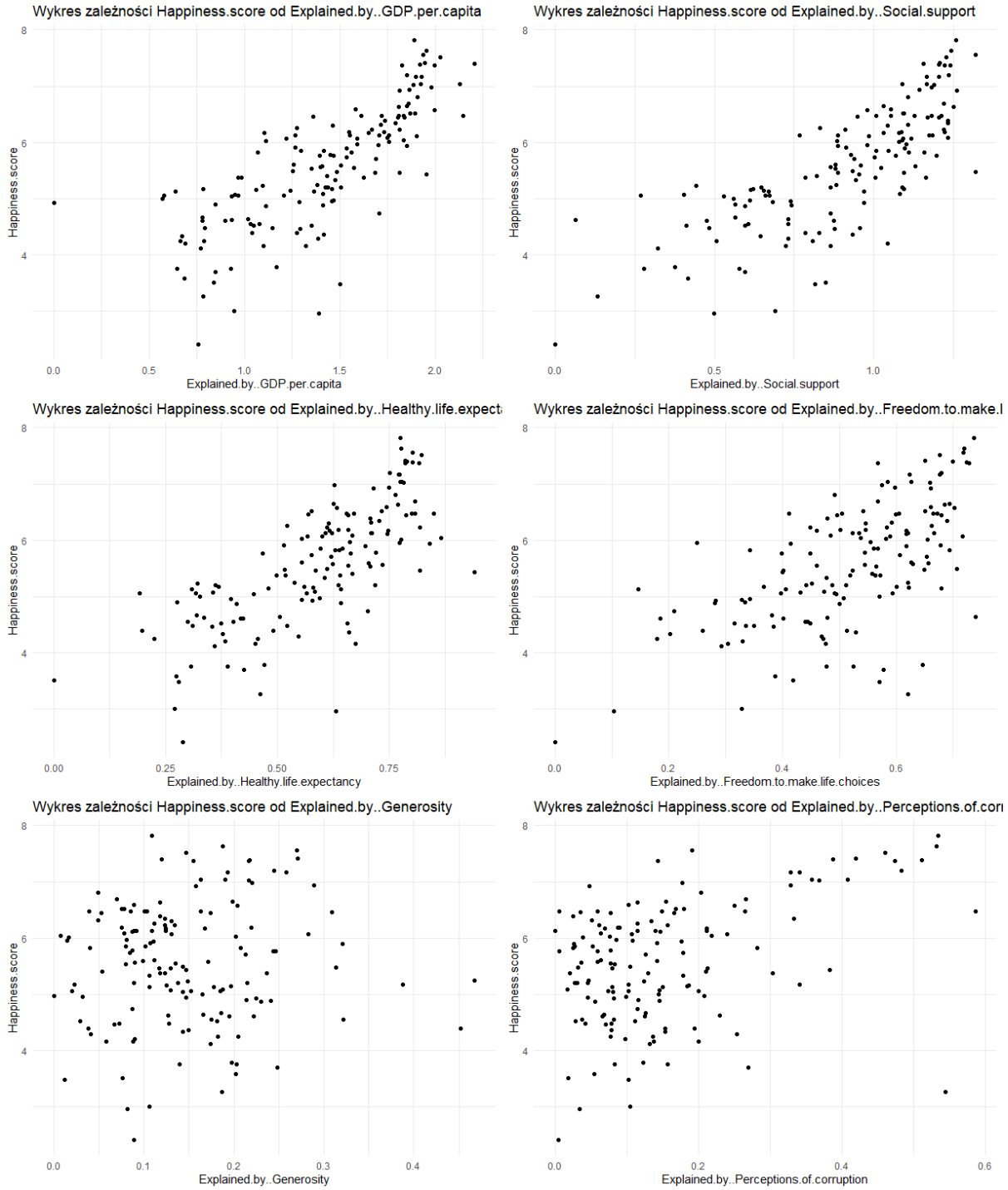
- Explained.by..GDP.per.capita
- Explained.by..Social.support
- Explained.by..Healthy.life.expectancy
- Explained.by..Freedom.to.make.life.choices
- Explained.by..Generosity
- Explained.by..Perceptions.of.corruption

Krok 2. Eksploracja i przygotowanie danych

Wprowadzone dane miały typ ‘character’. Aby zamienić wartości na numeryczne, przecinki, które występowały w wartościach zamieniliśmy na kropki. Kiedy dane miały już charakter numeryczny, przeszliśmy do usunięcia niepotrzebnych kolumn: Country, RANK, Whisker.high, Whisker.low oraz Dystopia..1.83….residual. Każdy kraj posiada odpowiedni poziom szczęścia społeczeństwa, tak więc kolumny Country oraz RANK nie wprowadzają do analizy ważnych informacji. Podobnie kolumny odpowiedzialne za przedziały ufności: Whisker.high oraz Whisker.low przedstawiają zakres, w którym mieści się prawdziwa wartość poziomu szczęścia i w tym przypadku również mogą zostać pominięte. Natomiast zmienna Dystopia..1.83….residual ma posłużyć jedynie jako punkt odniesienia. Na koniec usunęliśmy wartości “NA”, które mogły uniemożliwiać przeprowadzenie analizy. Finalna postać naszej bazy danych to 146 wierszy oraz 7 kolumn.

Otrzymaliśmy następujące wykresy zależności poziomu szczęścia od poszczególnych zmiennych objaśniających:

(są to kolejno: GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption)



Wstępnie możemy wnioskować, iż pierwsze cztery zmienne mogą wykazywać liniową zależność z objaśnianą zmienną Happiness.score. Wykres zależności od zmiennej Perceptions of corruption jest już bardziej rozproszony, z kolei w przypadku Generosity nie widać większej zależności.

Krok 3. Budowa modelu

Analizę rozpoczęliśmy od przetestowania modelu drzew klasyfikacyjnych. Jest to struktura drzewa, w której każdy węzeł reprezentuje test na wartości konkretnej cechy, każda krawędź wychodząca z węzła reprezentuje jedno z możliwych wyników tego testu, a każdy liść drzewa reprezentuje klasę lub wartość regresji.

Zmienna Happiness.score, którą analizujemy jest zmienną ciągłą, musimy ją więc zdyskretyzować. Po wyświetleniu summary(Happiness.score) wnioskujemy, że wartość szczęścia powyżej 6.3 możemy uznać za wysoką, wartości między 4.9 a 6.3 za średnią oraz wartości poniżej 4.9 uznajemy za niski poziom szczęścia. Badamy więc zdarzenie happiness.score <= 4.9 vs. 4.9 < happiness.score < 6.3 vs. happiness.score >= 6.3.

Tworząc zbiory treningowy i testowy, zbieramy do nich odpowiednio 100 oraz 46 wierszy. Po zbudowaniu drzewa, poprawność predykcji wynosi 0.67. Po przycięciu naszego drzewa do rozmiaru 4 wartość ta zmalała do 0.65, a po zmianie rozmiaru na 7, poprawność nie uległa zmianie i miała wartość 0.67.

Wynik, jaki otrzymaliśmy nie jest najgorszy, jednak po przetestowaniu modelu na różnych zbiorach testowych, ustawiając różne parametry set.seed, poprawność predykcji wahała się między wartościami 0.56 a 0.69, a drzewa w niektórych przypadkach były ciężkie do interpretacji. Dlatego kolejnym naszym krokiem było sprawdzenie, jak z naszymi danymi poradzi sobie model regresji liniowej.

Badania regresji rozpoczęliśmy od selekcji najlepszych podzbiorów zmiennych, na których będziemy opierać naszą analizę. Po utworzeniu zbioru treningowego, ponownie przeprowadziliśmy szukanie na nim najlepszych podzbiorów zmiennych. Najlepszymi modelami okazały się te z 5 i 6 zmiennymi. Na koniec postanowiliśmy jeszcze sprawdzić jak na naszych danych zachowa się modyfikacja regresji liniowej - regresja grzbietowa.

Krok 4. Ocena modelu

Dla najbardziej optymalnego parametru lambda wyszukanego walidacją krzyżową, otrzymaliśmy błąd średniokwadratowy na poziomie 0.24, oraz współczynnik R kwadrat na poziomie 0.81. Z racji, że ciężko było nam wybrać pomiędzy modelem regresji liniowej a regresji grzbietowej, oraz z racji na ewentualny problem wyboru ilości zmiennych w modelu regresji liniowej (w przypadku 6 zmiennych dostaliśmy niewiele mniejszy błąd, niż w przypadku 5), postanowiliśmy wykonać walidację krzyżową aby zobaczyć, jak nasze modele uogólniają się na nowe dane.

Przeprowadziliśmy walidację krzyżową z 10 podzbiorami i analizując wynik zdecydowaliśmy, że lepszym wyborem będzie model z 6 zmiennymi. Model regresji grzbietowej daje nieco lepszy współczynnik R kwadrat niż ten, uzyskany przy użyciu regresji liniowej. Analizując jednak fakt, że model regresji liniowej będzie się dobrze generalizował na nowe dane zdecydowaliśmy, że będzie on lepszym wyborem.

Ostatecznie zdecydowaliśmy się na model regresji liniowej z 6 zmiennymi. Są to: Social support, Freedom to make life choices, GDP per capita, Healthy life expectancy, Generosity oraz Perceptions of corruption. Średni błąd kwadratowy naszego modelu wynosi 0.0487, a współczynnik R kwadrat 0.77 i wnioskujemy, że wynik ten jest zadowalający.

Krok 5. Dopracowanie modelu

Podczas analizy sprawdziliśmy również, jakie wyniki dają inne modele oraz metody (takie jak regresja lasso, drzewa regresyjne). Ostatecznie jednak wybraliśmy ten o najbardziej wiarygodnych wynikach. Uzyskane predykcje są zadowalające, a model będzie dobrze generalizował się na nowe dane.

Kraków, 18.01.2024

Karolina Grzech, Krystian Bułat