

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 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 11 kolumn:

- Country - kraj, w którym przeprowadzono badania,
- Happiness.Rank - pozycja kraju w rankingu ogólnej oceny szczęścia. Wskazuje, jak dany kraj plasuje się w porównaniu do innych,
- Happiness.Score - wskaźnik ogólnego poziomu szczęścia w danym kraju,
- Dystopia - kraj teoretyczny, którego wartości są równe najniższemu narodowemu średniemu 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:

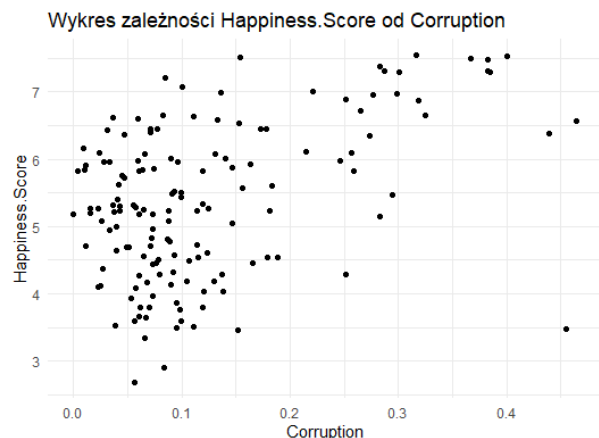
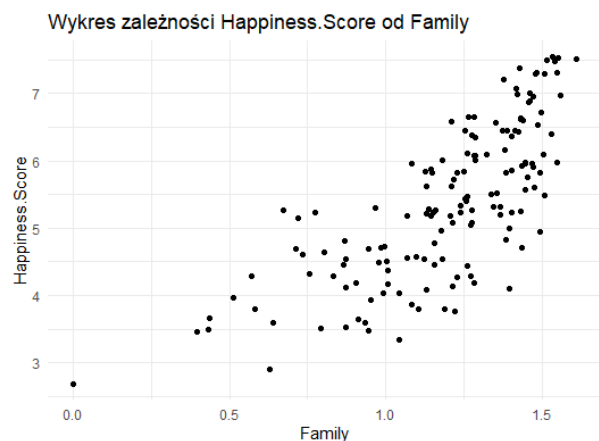
- Economy
- Family
- Health
- Freedom
- Generosity
- Corruption
- Job.Satisfaction

Krok 2. Eksploracja i przygotowanie danych

Wczytane dane musieliśmy oczyścić i przygotować do dalszej analizy. Zaczęliśmy od usunięcia niepotrzebnych kolumn: Country, Happiness.Rank oraz Dystopia. Każdy kraj posiada odpowiedni poziom szczęścia społeczeństwa, tak więc kolumny Country i Happiness.Rank nie wprowadzają do analizy ważnych informacji, natomiast zmienna Dystopia ma posłużyć jedynie jako punkt odniesienia. Usunęliśmy również ostatnią kolumnę, ponieważ nie było w niej danych. Na koniec usunęliśmy wartości "NA", które mogły uniemożliwić przeprowadzenie analizy. Finalna postać naszej bazy danych to 151 wierszy oraz 8 kolumn.

Otrzymaliśmy następujące wykresy zależności poziomu szczęścia od poszczególnych zmiennych objaśniających:

(są to kolejno: Economy, Family, Health, Freedom, Generosity, Corruption)



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

Krok 3. Budowa modelu

Nasza zmienna objaśniana była zmienną ciągłą więc analizę rozpoczęliśmy od

przetestowania modelu drzew regresyjnych. Po zbudowaniu modelu okazało się jednak, że dokładność takiego drzewa jest na poziomie 0.4, a model korzysta z 4 zmiennych. Spróbowaliśmy także sprawdzić czy przycięcie drzewa pozwoli uzyskać lepszą jakość predykcji. Najlepszy model jaki otrzymaliśmy zawierał 3 zmienne, a jego dokładność była na poziomie 0.41.

Przy okazji budowania drzew postanowiliśmy również sprawdzić jak z naszymi danymi poradzi sobie model drzewa klasyfikacyjnego. Zmienna Happiness.Score, którą analizujemy jest zmienną ciągłą, musimy ją więc zdyskretyzować. Po wyświetleniu summary(Happiness.Score) wnioskujemy, że wartość szczęścia powyżej 6.1 możemy uznać za wysoką, wartości między 4.5 a 6.1 za średnią oraz wartości poniżej 4.5 uznajemy za niski poziom szczęścia. Badamy więc zdarzenie $\text{happiness.score} \leq 4.5$ vs. $4.5 < \text{happiness.score} < 6.1$ vs. $\text{happiness.score} \geq 6.1$.

Tworząc zbiory treningowy i testowy, zbieramy do nich odpowiednio 100 oraz 51 wierszy. Po zbudowaniu drzewa, poprawność predykcji wynosi 0.67. Po przycięciu naszego drzewa do rozmiaru 4 wartość ta zmalała do 0.725, 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 0.25, oraz współczynnik R kwadrat na poziomie 0.84. Z racji, że ciężko było nam wybrać pomiędzy modelem regresji liniowej a regresji grzbietowej, oraz z uwagi 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: Economy, Family, Health, Freedom, Corruption oraz Job.Satisfaction. Średni błąd kwadratowy naszego modelu wynosi 0.22, a współczynnik R kwadrat 0.84 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

KG, KB