Sprawozdanie

Budowa modelu regresji liniowej
Semestr letni 2017/2018

Autor:
Artur Błażejewski, 200541

Politechnika Wrocławska
Wydział Elektroniki
Automatyka i Robotyka
ARR

1 Miary skuteczności modelu

Do porównywania jakości regresji liniowej wykorzystano błąd MSE (błąd średnio kwadratowy) oraz współczynnik determinacji R^2 (najlepsza wartość 1.0). W obliczaniu skuteczności modelu zostanie wykorzystana walidacja krzyżowa z pięcioma iteracjami.

2 Metoda

Zastosowano metodę LinearRegresion z biblioteki sklearn z pakietu linear_model.

3 Przygotowanie danych

Z danych usunięto wiersze, w których brakowało rekordów lub posiadały one błędną formę.

Przed uczeniem modelu dane są zawsze przetasowane w niezmienny sposób oraz podzielone na dwa zbiory:

- Zbiór treningowy - 80% całego zbioru danych
- Zbiór testowy

W celu oczyszczenia danych z próbek niepozwalających na poprawne przewidywanie wykorzystano wiedzę zawartą w artykule[1] gdzie sugerowano, że cena niewolnika stanowi wielomian szóstego stopnia.

3.1 Oczyszczanie danych

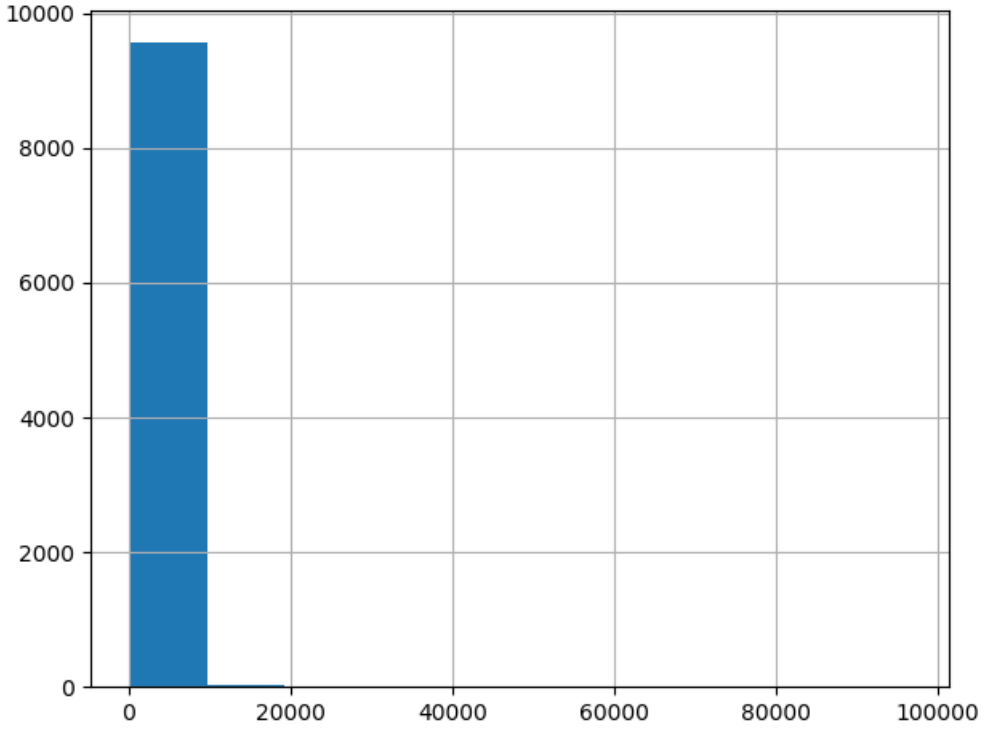
Usunięto	MSE	R^2
Nic	-5134398	-0.08
Cena powyżej 2000$ 3.1.1	-113938	0.29
Sprzedaż powyżej jednej osoby 3.1.2	-96511	0.33

3.1.1 Cena

Histogram 1 pokazuje rozkład ceny bez edycji, po usunięciu rekordów powyżej 10 tys. otrzymujemy 3 a gdy usunięto wartości powyżej 2 tys. **??** i dopiero ten rozkład cen przypomina normalny dlatego zdecydowano się na przedział do 2 tys.

3.1.2 Sprzedaż w zestawach

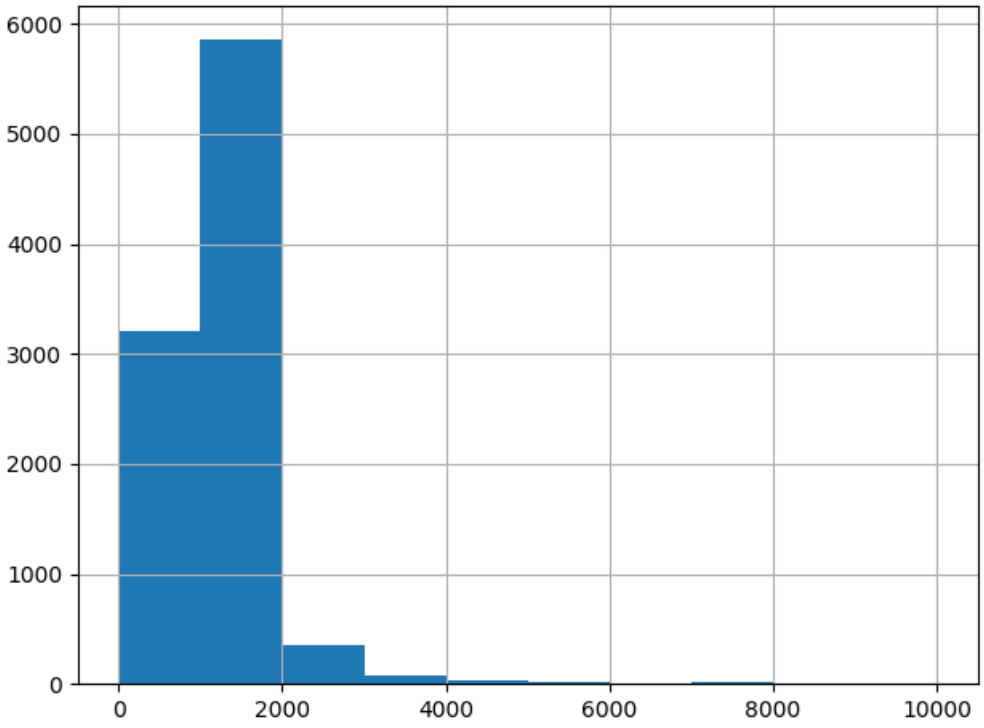
Usunięto wszystkie transakcje dotyczące więcej niż jednej osoby, ponieważ wpływają one negatywnie na wynik przewidywania. W pakiecie są sprzedawani niewolnicy w różnym wieku, a cena jest podana jedna dla wszystkich. Uśrednienie ceny nie przyniosłoby pozytywnych rezultatów.



Rysunek 1: histogram cen

4 Analiza błędów

Na podstawie krzywej uczenia pokazanej na rysunku 4 należy stwierdzić, że model posiada wysokie odchylenie, co może oznaczać, że cierpi na niedouczenie i pożądane jest wprowadzenie nowych cech.

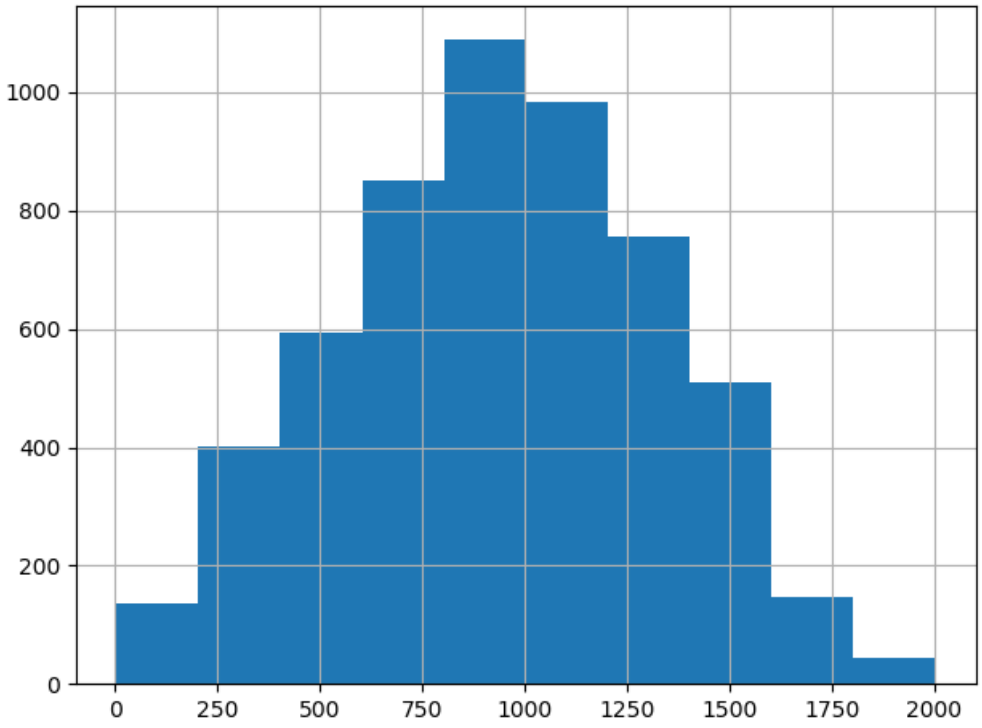


Rysunek 2: histogram cen w przedziale do 10 tys.

5 Poprawa modelu poprzez dodania nowych cech

5.1 Wyniki przewidywania

Cecha	MSE	R^2
Płeć 5.2.2	-93751	0.35
Płeć/wiek 5.2.3	-91862	0.37
Data sprzedaży 5.2.4	-82296	0.43
Pochodzenie kupującego 5.2.5	-81343	0.43
Pochodzenie sprzedającego 5.2.6	-81267	0.43
Kolor skóry 5.2.7	-81116	0.44
Kolor skóry modyfikacja 5.2.8	-81009	0.44



Rysunek 3: histogram cen w przedziale do 2 tys.

5.2 Cechy modelu

5.2.1 Wiek

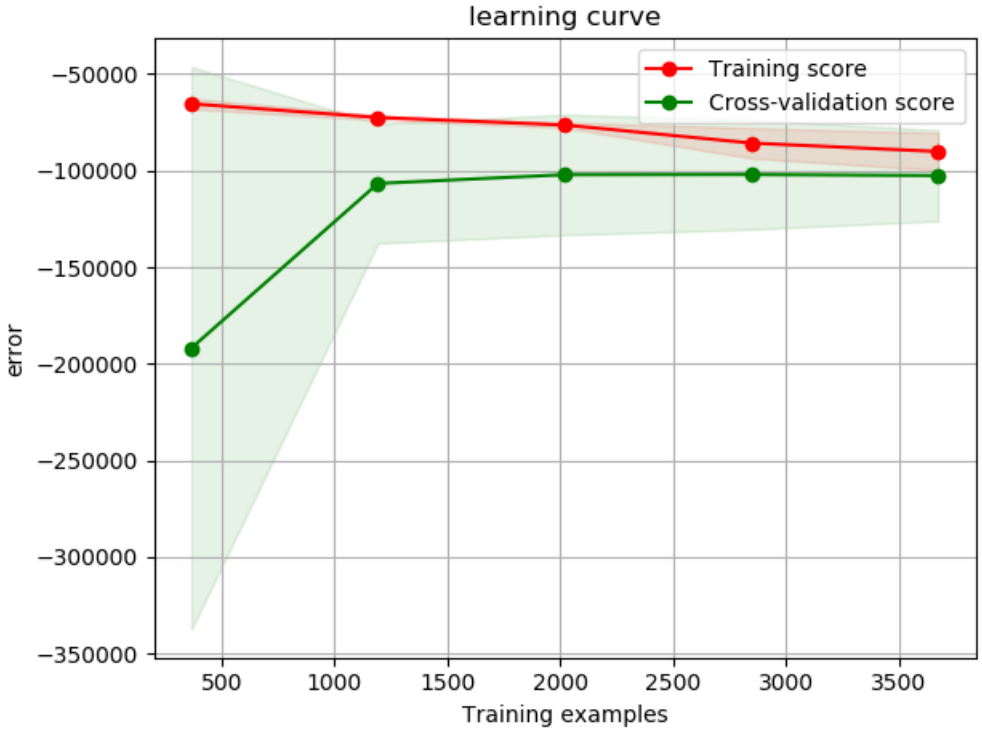
Poddano analizie wpływ wartości wielomianu na wielkość błędu i potwierdzono, że wielomian szóstego stopnia jest optymalny, ponieważ kolejne nie przynosiły znacznej poprawy (poniżej 1%).

5.2.2 Płeć

Ponieważ w zbiorze płeć jest oznaczona jako M oraz F wartości te zmieniono na 1 dla mężczyzn, oraz 0 dla kobiet. Tak sformułowaną cechę dodano do modelu.

5.2.3 Wiek oraz płeć

Ponieważ najwyższa cena mężczyzn i kobiet przypada na różny wiek, została stworzona cecha, która dla kobiet wynosiła 0, a dla mężczyzn stanowiła



Rysunek 4: Wykres błędu MSE dla zbioru treningowego i walidacyjnego w zależności od wielkości zbioru treningowego

wiek. Ponieważ wiek jest funkcją wielomianową, rozważono wielomiany stopnia 1,2,3,4,5,6 i ostatecznie wybrano wielomian stopnia 3.

5.2.4 Data sprzedaży

Ceny niewolników różniły się znacznie w zależności od pory roku, ponieważ w okresie letnim w Nowym Orleanie panowała żółta febra, mało osób decydowało się na zakup w tym czasie. Ponieważ okres ten obfitował w różne wydarzenia historyczne mające wpływ na cenę, również rok niesie pewne pozytywne informacje w procesie przewidywania.

Ponieważ przewidywanie odbywa się w okresie od października 1956 do sierpnia 1861 cecha jest opisana za pomocą numeru miesiąca zakupu obliczanego wg formuły 1

$$cecha = miesiac + 12 * (rok - 1956) \tag{1}$$

5.2.5 Pochodzenie kupującego

Cecha ma wartość 1, gdy kupujący jest z Nowego Orleanu, w przeciwnym wypadku 0.

5.2.6 Pochodzenie sprzedającego

Cecha ma wartość 1, gdy sprzedający jest z Nowego Orleanu, w przeciwnym wypadku 0.

5.2.7 Kolor skóry

Kolejną cechę stanowi kolor skóry, został zapisany wg wzoru:

- negro = 2
- mulatto = 3
- griff oraz light griff = 4
- yellow = 5
- pozostałe kolory oraz brak koloru 1

5.2.8 Kolor skóry tylko kobiety

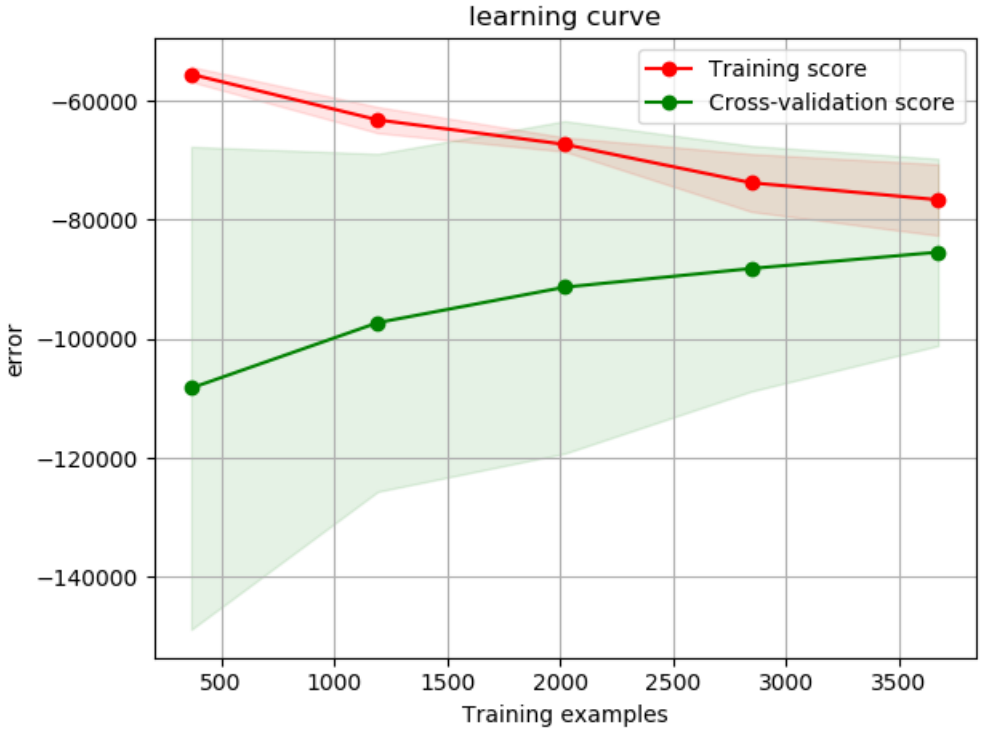
Wg autorów artykułu[1] kolor skóry ma znacznie większy wpływ na cenę sprzedawanych kobiet. Zmodyfikowano wcześniejszą cechę kolor, ustawiając wartość 0 dla mężczyzn.

6 Podsumowanie

6.1 Wybór cech

Na podstawie krzywej uczenia przedstawionej na rysunku 5 wyznaczonej dla modelu wyszkolonego dla wszystkich nowo wybranych cech można stwierdzić, że dostarczenie większej ilości danych powinno spowodować poprawę modelu.

Najlepszy wynik to MSE = 81009, $R^2 = 0.44$ oraz 73 średni błąd 73 (MSE przez wielkość zbioru walidacyjnego).



Rysunek 5: Wykres błędu MSE dla zbioru treningowego i walidacyjnego w zależności od wielkości zbioru treningowego dla nowo wybranych cech

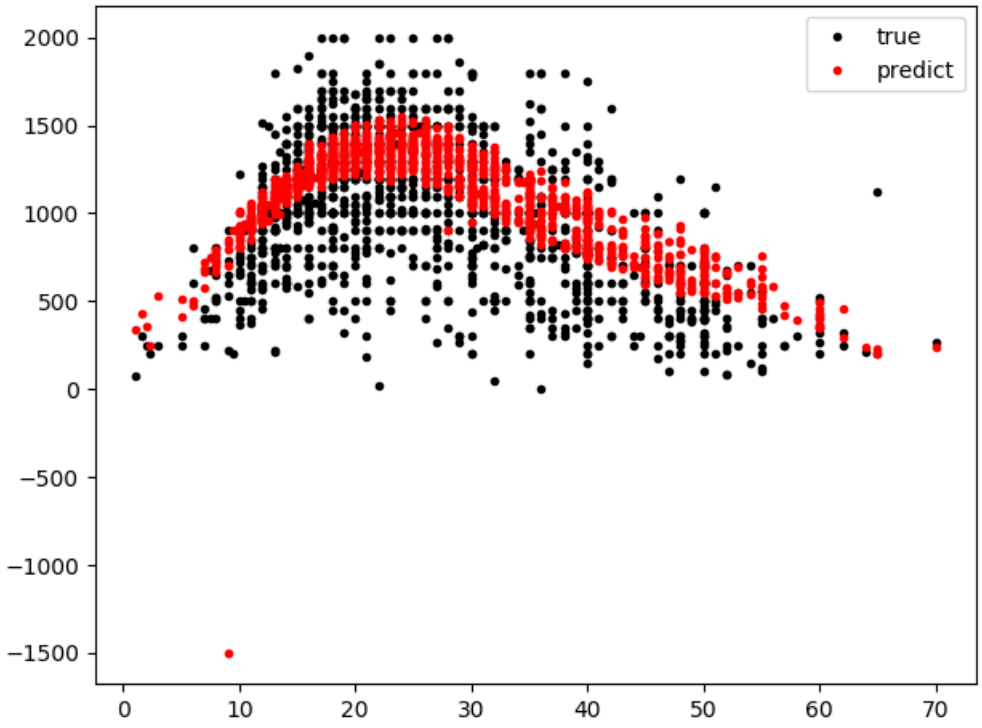
6.2 Wyniki dla zbioru testowego

Na wcześniej wydzielonym zbiorze testowym otrzymano błąd MSE=115761, R^2=0.29 oraz 84 (MSE przez wielkość zbioru walidacyjnego). Na wykresie 6 można zobaczyć rozkład prawdziwych i przewidywanych wartości cen. Przewidywanie ceny niewolnika wymaga wielu parametrów i jest obarczone dużym błędem, należy się spodziewać, że znaczny wpływ na cenę miały niepodane dane jak wzrost, waga. Otrzymana wartość MSE względem ilości prób testowych pokazuje wzrost błędu dla przypadku testowego względem walidacyjnego. Oznacza to, że model ma pewne problemy z generalizowaniem zagadnienia pomimo wykorzystania zbioru walidacyjnego.

7 Uruchomienie

7.1 Pliki

Do raportu zostały dostarczone następujące pliki:



Rysunek 6: Przewidywane i prawdziwe wartości ceny w zależności od wieku dla zbioru testowego

- LoadDataSet - zawiera funkcje do pobrania danych oczyszczenia z błędów, dodaje i modyfikuje cechy na potrzeby uczenia modelu
- LearnModel - pozwala na uzyskanie wartości błędów oraz wyrysowanie krzywej uczenia, plik stworzony na potrzeby dobrania najlepszych parametrów oraz cech modelu
- ModelTest - w pliku znajduje się skrypt odpowiedzialny za wyuczenie modelu a następnie przetestowanie go na wydzielonym zestawie testowym

W plikach LearnModel oraz ModelTest znajduje się zmienna path które zawiera ścieżkę do pliku z danymi (CalomirisPritchett_data.xlsx) plik pobrany z strony dotyczącej artykułu[1]

Skrypty zostały napisane w języku Python w wersji 3.6.

Literatura

[1] *Betting on Secession: Quantifying Political Events Surrounding Slavery and the Civil War, Charles W. Calomiris and Jonathan Pritchett,* `https://www.aeaweb.org/articles?id=10.1257/aer.20131483`