Przewidywanie ceny samochodu

Denys Troshchylo

12 czerwca 2023

Spis treści

1 Definicja problemu badawczego

W tym projekcie badawczym celem jest stworzenie modelu, który bedzie przewidywać ceny samochodów na podstawie dostepnych danych. Celem jest stworzenie modelu, który bedzie dokładnie przewidywał ceny na podstawie różnych atrybutów samochodów. Cały projekt jest napisany w jezyku python z wykorzystaniem różnych bibliotek

2 Zdobycie danych

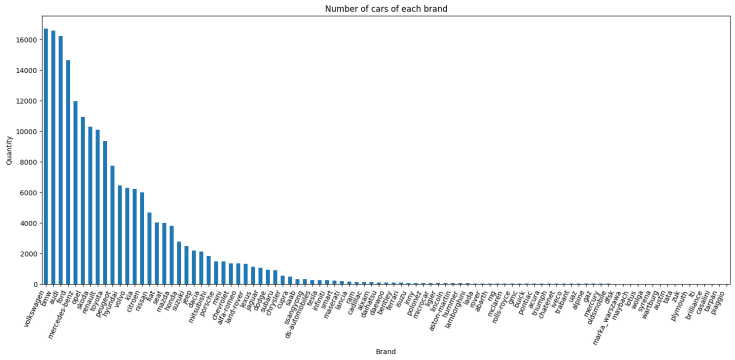
Do realizacji pobierania danych dla tego projektu użyłem biblioteki *scrapy*. Dane pobierałem ze strony `https://www.otomoto.pl/`. Podczas implementacji danej cześci napotkałem kilka problemów. W pierwszej wersji tego skryptu otwierałem każda strone z ofertami samochodów, a nastepnie otwierałem każda strone każdego samochodu po to żeby pobierać wiecej atrybutów dla każdego samochodu. W tym przypadku miałem problem z tym, że skrypt działal tak szybko że strona z której pobierałem dane blokowała mnie po 5 minutach skrapowania. Żeby wyeliminować ten problem zrobiłem losowe opóźnienie przed każdym pobieraniem. Potem wystapił taki problem, że po pobraniu mniej wiecej 15 tysiecy stron predkość pobierania nagle spadała i nic sie nie zapisywało do bazy samochodów, później sie okazało, że tak naprawde ta strona nie miała 6 tysiecy stron z ofertami, a tylko 500 i potem jak skrypt próbował wejść np. na 501 strone to zostawał przekierowany na 500. Wiec zmieniłem strategie i pobrałem dla każdej marki samochodu dostepne modele i przechodziłem już po stronach modeli i nie wchodzac na strone z samochodem skrapowałem dane tego samochodu.

3 Przygotowanie, wyczyszczenie danych

Przed przystapieniem do analizy danych, przeprowadziłem kilka kroków w celu przygotowania i wyczyszczenia danych. Najpierw to było sprawdznie czy mam wszystkie atrybuty dla każdego samochodu jak sie okazało to wystepowały takie oferty w których nie było przebiegu, to oznacza że samochód jest nowy, wiec po prostu wstawiałem 0 w pole przebiegu. Również były takie oferty które nie miały pojemności skokowej, ale to wystepowało wyłacznie u samochodów które miały rodzaj paliwa Elektryczny lub Hybryda, w tym przypadku, jak i w poprzednim, zamieniałem na 0. W sumie udało mi sie pobrać 195000 opisów aut. Kolejny problem był taki, że niektóre modele miały dużo stron i jak sie przechodziło przez te strony czasami wystepowały powtórzenia ofert, żeby poradzić sobie z tym problemem usunalem duplikaty. Po wyeleminowaniu tych problemów zostało +-100 samochodów które zawierały jakieś błedna dane, dlatego usunałem je. Niektóre ceny były w EUR, wiec po prostu rzeliczyłem walute z euro na złotówki. Na wyjściu z tego dostałem 192000 poprawnych jednostek z danymi.

4 Wstepna analiza danych

Przeprowadziłem wstepna analize danych w celu zrozumienia zawartości zbioru i identyfikacji ewentualnych zależności. Wykorzystałem techniki wizualizacji danych za pomoca biblioteki *matplotlib*. Przykładowo, wygenerowałem wykres przedstawiajacy stosunek ilości pobranych aut do marek. Inne wykresy do analizy danego zbioru umieściłem w Jupyter Notebook'u



Rysunek 1: Stosunek ilości pobranych aut do marek

5 Przygotowanie modelu

Przygotowanie modelu składało sie z kilku etapów. Wybrałem sobie model drzewa decyzyjnego z biblioteki *sklearn*. Już gotowy sformatowany zbiór danych umieściełem w DataFrame z biblioteki *pandas*, żeby móc wytrenować model. Nastepnie wytrenowałem model oraz zapisałem go do pliku żeby nie robić tej samej czynności wiele razy.

6 Wnioski, komentarze, ulepszenia

Po przeprowadzeniu analizy danych i wytrenowaniu modelu, doszedłem do nastepujacych wniosków. Mój model osiagnał dokładność na poziomie 10-30 procent i uważam, że dla takiego, dość małego, zbioru atrybutów to jest dobry rezultat. Najtrudniajsza cześcia tego projektu dla mnie było skrapowanie danych bo było wiele niuansów które trzeba uwzglednić. Żeby polepszyć precyzyjność tego programu, moim zdaniem, można by było:

1. Pobierać wiecej atrybutów wchodzac na strone każdego auta.
2. Pobrać dane również z innych stron które sprzedaja samochody.
3. (Zaawansowane) Pobierać zdjecia samochodów oraz wykrywać widoczne na nich uszkodzenia.

7 Link do projektu

Projekt jest dostepny na platformie GitHub pod adresem:
`https://github.com/troshchylo/predicting-car-prices`