Prices Forecasting of Cars Sales Advertisements

Motivation

Understanding the principles of cars prices formation is important for customers, dealers and intermediaries. If you are buying a brand new car from an authorized dealer, you face sticky prices. However, if you want to buy a used car from an individual or unauthorized dealer, prices can vary sharply. If you want to sell your car, you probably do not know the exact price of your vehicle on aftermarket. Car price obviously depends on broad variety of different parameters and it is very hard to keep all in mind if you are not a car reseller. Ability to forecast the car price based on its characteristics could save a lot of money for customers and increase profits of dealers and intermediaries.

In addition, I have personal motivation in this project because I plan to buy a car in the nearest future.

Project stages

1. **Data searching.** First of all, I limited variety of cars for analysis to most popular D-segment cars in Russia[1]. https://moscow.auto.ru/ – one of the most visited[2] automotive sites in the Russian internet and it was selected as data source for analysis. This website has most number of cars selling advertisements of new and used cars. One should remember that this would be supply prices. Without loss of generality we can make an assumption that supply price exceeds actual price at a certain percentage because it is very common to have 5%-10% discount from car seller (used car usually has a number of shortcomings). All materials for this project can be found at https://github.com/BiXiC/auto_ru
2. **Data collection.** To scrap webpages with cars advertisements, I used Python 3 and BeautifulSoup library, developed scrip that parsed all advertisements from Moscow and +200 km area around it. Final dataset has 7715 advertisements of D-segment cars. Car advertisement contains information about car price, seller, owner, list of car characteristics like engine capacity, horse power, gear box type, fuel type, color, age, run, warranty, Views of advertisement, number of owners, and various car options which were mentioned in advertisement (like, for example, xenon headlights, ABS or leather upholstery).

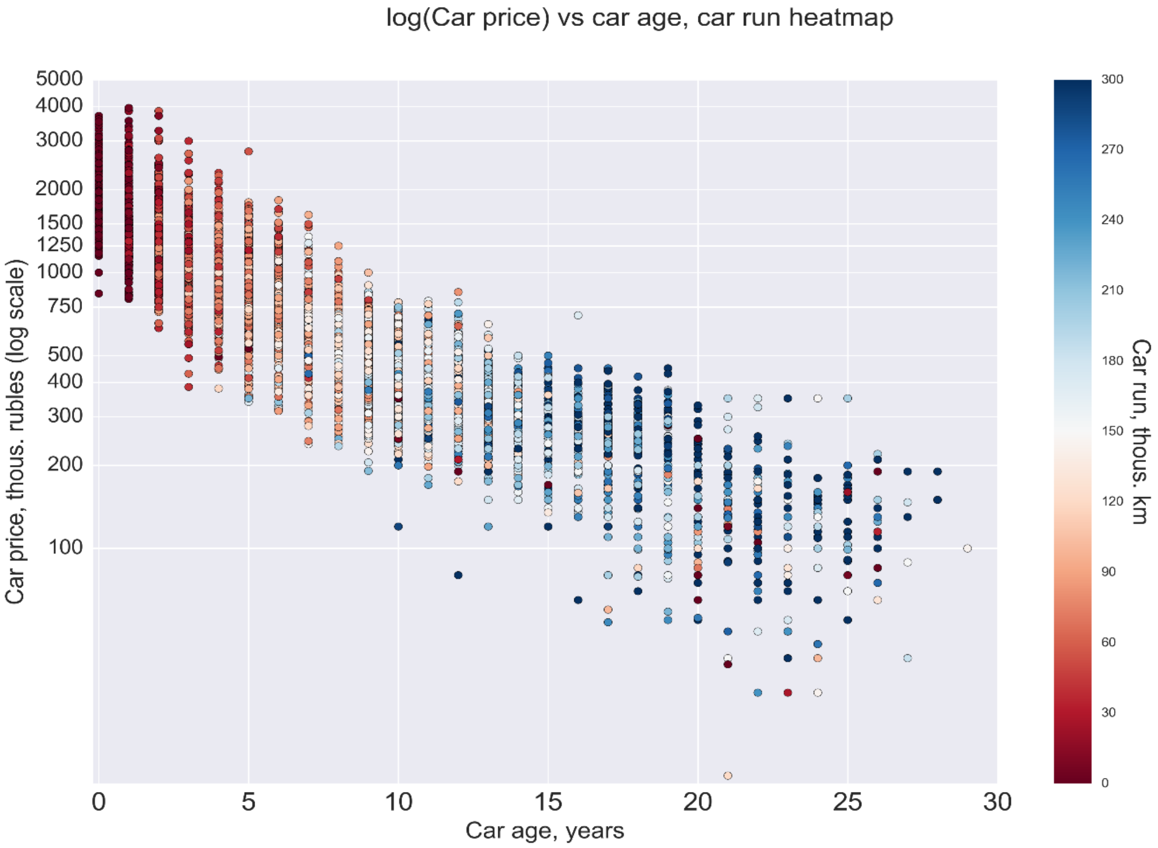
[1] Short list of models for analysis: Kia Optima, Honda Accord, Hyundai i40, Hyundai Sonata, Toyota Camry, Mazda 6, BMW 5er, Audi A6, Ford Mondeo, Infiniti G35, Nissan Teana, Opel Insignia, Volvo S60

[2] According to TNS Web Index Report in February 2016, auto.ru has 10.8% monthly reach, Moscow population and 5.7% monthly reach for Russian population 12-64 years old.

Main problem here could occur when we would decide to add another car advertising platform. It would be hard to find all advertisements duplicates from different platforms.

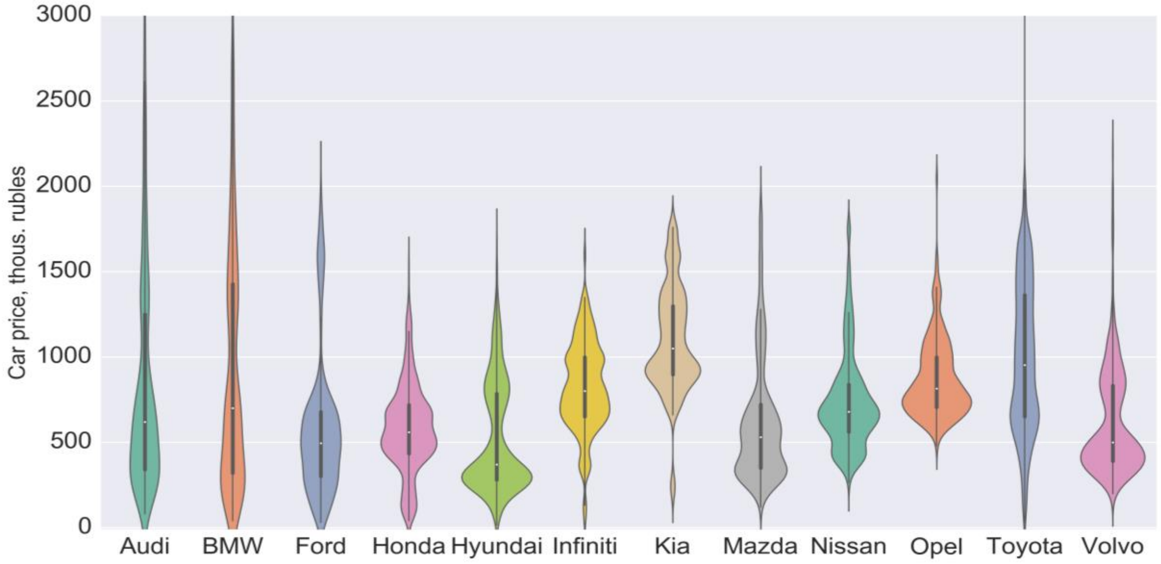
3. Exploratory analysis.

It is common sense that car price decreases on certain percentage every year. So as it becomes older price decreases for smaller amount of money. So obvious form of relation for price vs age would be log-linear. Plot 1 represent relation between log of car price and car age. Heat map depicts car run. As we can see on this plot, our hypothesis is confirmed. Car run also has dependence on price.



Plot 1

Plot 2 depicts cars brands densities of price. We can see that most of cars prices are below 1 mio rub but Kia brand seems to be the most expensive. This is not natural and the reason is I picked up only Optima model from Kia wich is availible only from 2011 while other models have pretty long history and market is saturated with old cars.



Plot 2

4. **Forecasting**

   (i) *Linear regression methods*

   To get forecasting benchmarks I fit two simple linear models (OLS) based on results of exploratory analysis. First model uses short list of basic car features. Second has a bit more extended list. I used MAPE to estimate error. On test data (20% of data, stratified to car name + generation variable) first model got MAPE 10.6% and second – 9.91%. This is not bad result for simple linear regression and can be obtained very fast.

   (ii) *Tree based methods*

   a. Gradient Boosting Regressor model on same train/test split got MAPE 9.14% on test data. (3000 trees, max_depth = 3)

   b. XGboost model on same train/test split got MAPE 9.05% on test data (num. round = 1000)

   Based on MAPE metric ML models could easily improve predictive ability of linear model and well-tuned XGboost method should significantly decrease an error. Obviously cross validation should be done for precise estimation of out of sample error on validation set but due to the lack of time it was left for future analysis.

5. **Further development.** On the next step of this project accurate and neat ML model should be developed. Cars models list should be extended and to all automotive segments. Various features from text comment from seller should also increase predictive ability. Other car-selling platform can be added. After that advisory website can be created and it can work as a service for customers for predicting optimal price for their vehicle. Another option for this model is integration into automotive website.