Capstone 2 Milestone Report 1

Problem Statement

Used car market has always been frequently discussed as the prices and depreciation of used cars are very tricky for both sellers and buyers and there is a considerable amount of transaction cost. This project is dedicated to provide some insights on the pricing mechanism and help to reduce the information cost.

For both dealers/sellers and buyers, this model would provide a good reference on pricing the vehicle and price transparency, which could improve the market efficiency by reducing the information cost and simplify the decision making process.

Data Source

https://www.kaggle.com/lepchenkov/usedcarscatalog/data

Author: Kirill Lepchenkov

The dataset is collected from various web resources in order to explore the used cars market and try to build a model that effectively predicts the price of the car based on its parameters (both numerical and categorical)

The data is scraped in Belarus (western Europe) on the 2nd of December 2019.

Data Cleaning

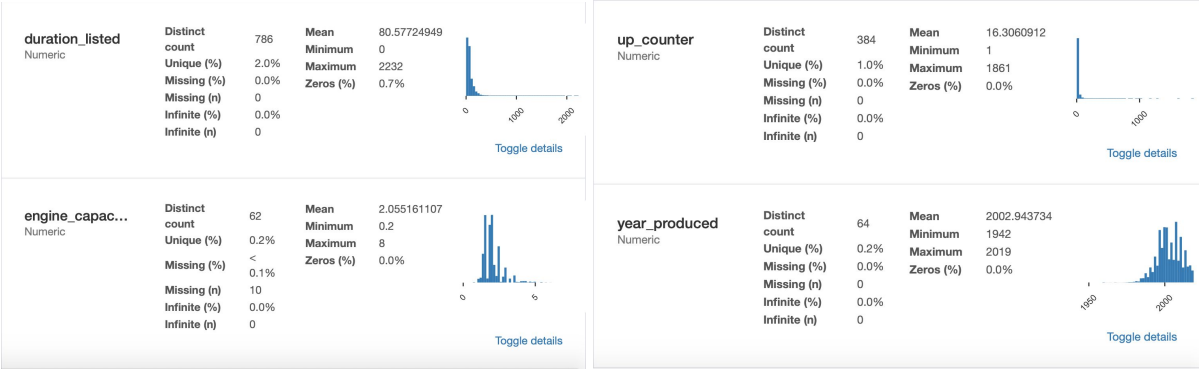
The original dataset has 38531 entries and 30 columns, all of them are non-null but 10 of them are nan and 2 columns are float64, 18 columns are int64, and 10 columns are object.

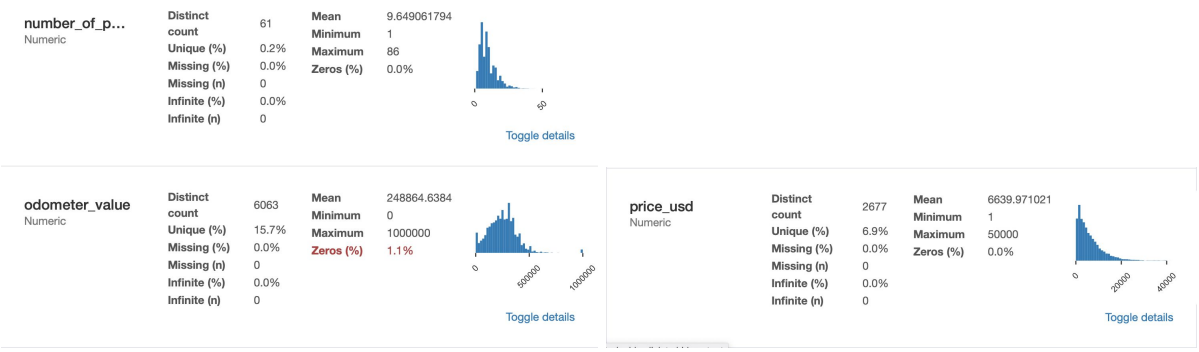
Data Wrangling

Boolean columns were then converted to numerical values. For categorical columns with less than 10 distinct values, one hot encoding was applied and then high correlation columns were dropped. For categorical columns with high cardinality, target encoding and hashing encoding were applied to generate 2 versions of the data, and missing values were filled with mean after splitting training and testing data.

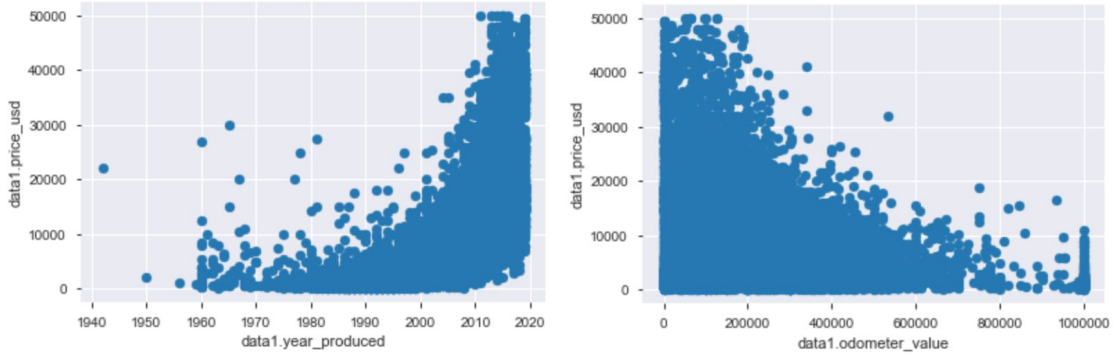
Exploratory Analysis

Some columns such as price and odometer value seem to have distribution patterns.

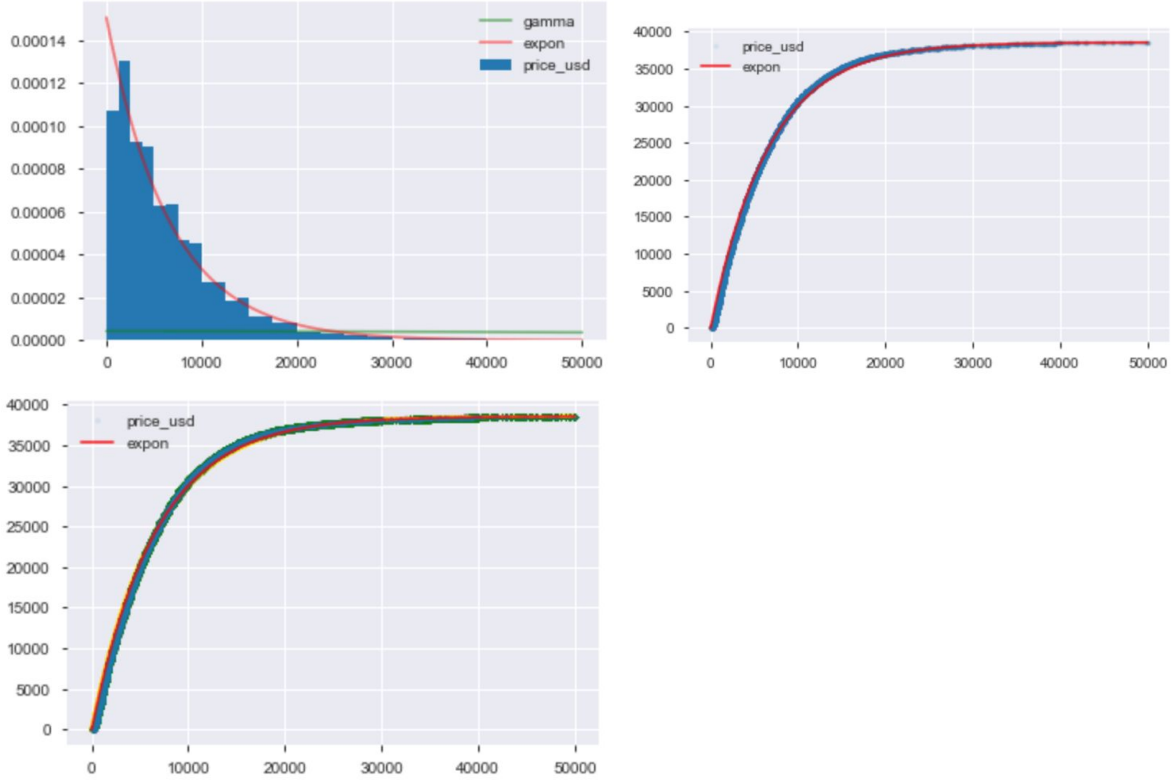




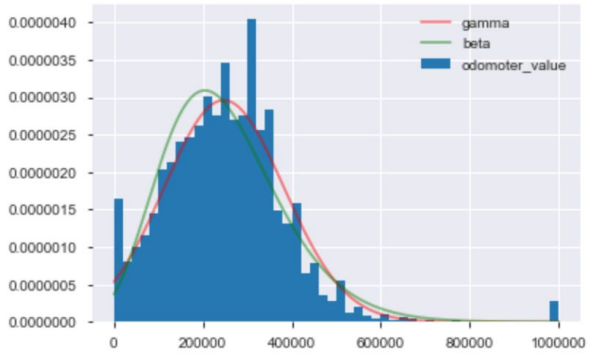
And some columns seem to be correlated: year produced and price seem to be positively correlated while odometer value and price seem to be negatively correlated, which is aligned with common sense on used cars.

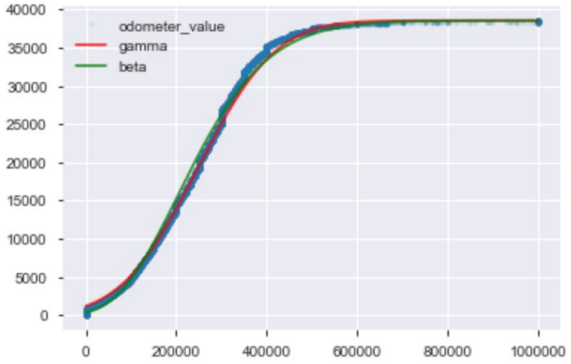


From price's histogram and empirical cumulative distribution it looks like price is exponentially distributed, and the bootstrap test (n=1000) suggests so as well.



As for odometer value's histogram and empirical cumulative distribution, it seems to be gamma distributed.





And year produced's histogram, empirical cumulative distribution and the bootstrap test (n=1000) suggest it is log gamma distributed.