Анализ и прогнозирование цен на поддержанные авто в США

Лащёнов Максим

*Технический университет Молдовы*

e-mail: [maxim.lascionov@iis.utm.md](mailto:maxim.lascionov@iis.utm.md)

**Абстракт**

В данной статье представлен анализ данных о подержанных авто на основе набора данных. Был разведочный анализ данных, графическое представление, определения основных переменных, а также построена модель для определения ценообразования на автомобили.

**Введение**

Данное исследование не имеет какой-либо конкретной цели, а скорее личный интерес, для анализа рынка поддержанных автомобилей США, так как по своей сути он является самым большим в мире, а следовательно большинство подержанных автомобилей из США попадает в страны Европы и СНГ. Поэтому лично мне было достаточно интересно было просто сделать анализ на основе различных гипотез, даже не связанных с ценообразованием.

**Автомобиль** — моторное безрельсовое дорожное и/или внедорожное, чаще всего автономное,транспортное , используемое для перевозки и/или грузов, имеющее от четырёх колёс.

Основное назначение автомобиля заключается в совершении транспортной работы. Автомобильный транспорт в промышленно развитых странах занимает ведущее место по сравнению с другими видами транспорта по объёму перевозок пассажиров. Современный автомобиль состоит из 15—20 тысяч деталей, из которых 150—300 являются наиболее важными и требующими наибольших затрат в эксплуатации и обслуживании.

Авторынок б/у автомобилей в США представляет собой значительную исследовательскую область в контексте мировой автомобильной индустрии. Его огромный объем и разнообразие обусловлены несколькими важными факторами, которые подлежат глубокому анализу.

Популяция и обширные территории Соединенных Штатов, а также соответствующая инфраструктура, играют значительную роль в формировании спроса на автомобили.

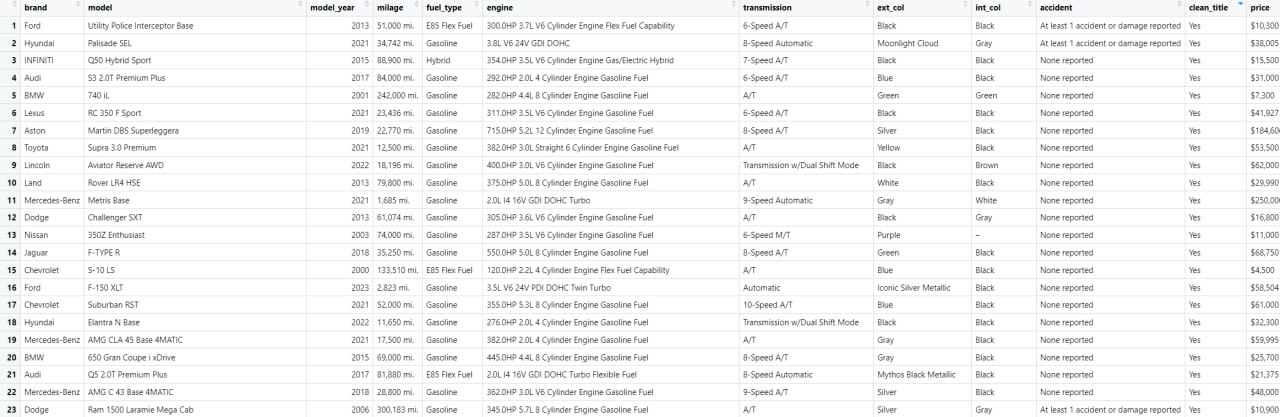
Экономическая стабильность и высокий уровень доходов создают благоприятную среду для покупки и обслуживания транспортных средств. Историческая перспектива также играет ключевую роль в структуре этого рынка, учитывая развитие автомобильной промышленности в США на протяжении десятилетий.

**Материалы и методы**

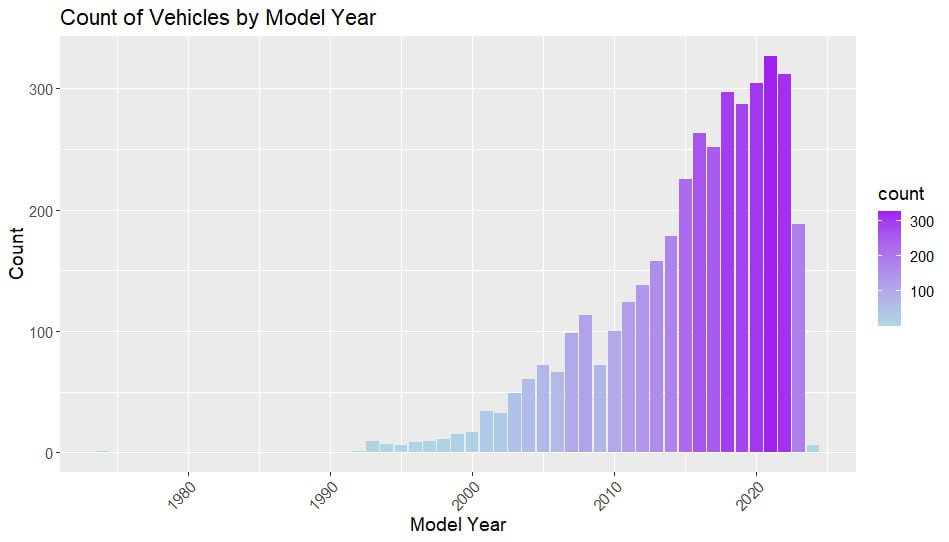
По своей сути, как материал я использовал открытый набор данных, взятый с kaggle.com. В нем указано 4000 позиций, на различные машины. Из переменных,которые я взял за основу, как главная в моей гипотезе, был пробег и год выпуска автомобиля, позже при построенний модели регрессий, взгляд на ценнообразование изменился, но это позже. В этом наборе данных, было 11 переменных, 10 из которых были независимы, и 1 зависимая (цена). На рисунке 1, указан пример нескольких строк данных из набора данных.

Для анализа набора данных использовался язык программирования R, библиотеки ggplot для построения графиков, dplyr, для дополнительных функций для этого языка. Пример графиков на рисунке 2 и 3.

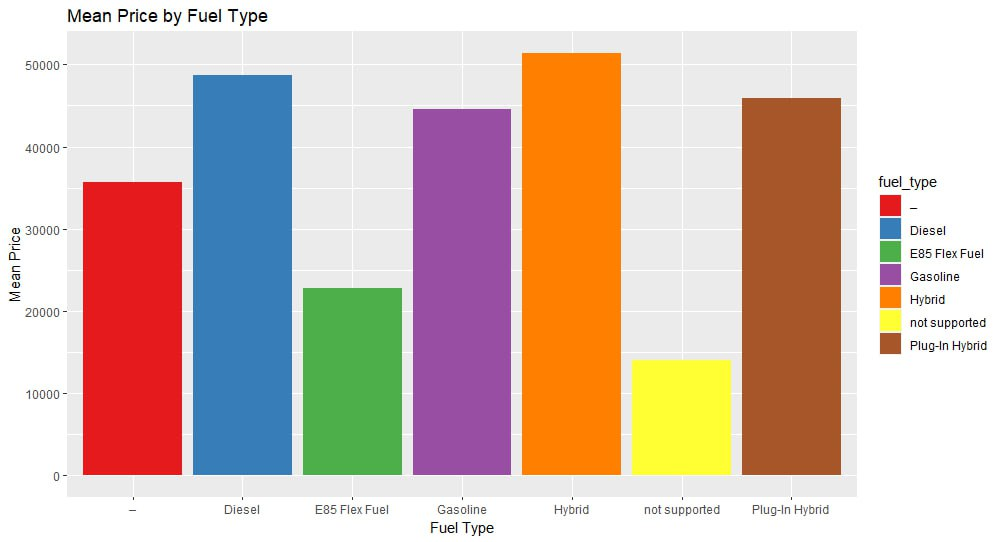
Для построение модели я использовал линейную регрессию и только ее, логистическую модель строить смысла не было, так как в набор отсутствует поля со значениями 1 и 0. Один из примеров построенной мной модели и значений которые она выдавала на рисунке 4.



**Рисунок 1**

****

**Рисунок 2**

****

**Рисунок 3**

Из обработки данных, были использованы стандартные методы, перевод нескольких строковых полей в числовые, а также избавления от пустых полей.

**Результаты**

В ходе анализа данных я получил достаточно интересные результаты по своим гипотезам. Во-первых, на рынке США до сих пор большее количество машин которые ездят на бензине, в то время как у нас в Молдове большинство автомобилей на дизеле. Во-вторых, это то что большинство машин которые находятся в продаже, они в период с 2013 по 2023 года выпуска.

Насчет модели линейной регрессии, из результатов что я получил, отметая менее значимые переменные, лучший результат который я смогу получить это:  
 R-squared: 0.771

p-value < 2.2 \* 10^-16

F-statistic: 2.011 on 1254 and 749 DF

Residual standard error: 78140 on 749 degrees of freedom

Данные результаты были получены на основе построения модели из переменных: модель, пробег, бренд, дата выпуска, информации о ДТП, чистый статус.

**Обсуждение**

В ходе анализа, а особенно построения модели, на основе данных, я опроверг свою гипотезу ценообразования автомобилей. Свою изначальную гипотезу я строил по принципу, что больше всех на цену влияет пробег авто, а также год его выпуска. но на самом деле, после того как я построил две модели линейной регрессии я опроверг свою гипотезу.

Первая модель была построена на основе переменных: двигатель, трансмиссия, цвет кузова, информация о ДТП и цвет внутренней отделки.

Вторая модель: модель, пробег, бренд, дата выпуска, информации о ДТП, чистый статус.

В первой модели я получил результат R-squared около 0.65, а вот во втором случае 0.771. Это меня удивило, так как обе эти модели, говорят, о том что пробег никак не влияет на ценообразование. Больше всех, не считая бренд и модель, влияет на саму машину цвет внутренней отделки автомобиля, а на втором месте цвет кузова.

Это меня удивило, но после я понял, проведя анализ, так как в США крайне большой выбор машин и количество людей, людям важнее то как выглядит машина и какой у нее салон, нежели ее качество, пробег и год выпуска, к тому же так как люди в США в основном передвигаются на автомобилях, исключая несколько штатов, пробег в этих местах не учитывается в образовании цены на автомобиль.

**Выводы**

Данное исследование показало, насколько обширен рынок б\у автомобилей, и насколько трудно определять цену, без фото самих автомобилей. Данное исследование,могло бы на самом деле хорошо послужить бизнесам, которые связанные с перепродажей автомобилей. Но для этого данный датасет, нужно дополнять фотографиями самих авто, дабы оценить состояние кузова.

**Библиография**

1. Коды и source-датасета (<https://github.com/h0pefu1/used-cars-analyze>)
2. Построенние модели линейной регрессии в языке R (<https://www.datacamp.com/tutorial/linear-regression-R>)
3. GGPlot построенние график (<https://ggplot2.tidyverse.org/>)
4. Источник данных для датасета (<https://www.cars.com/>)
5. Анализ данных основа и терминология

(https://habr.com/ru/articles/352812/)