Анализ стоимости автомобиля

ML DS Case

Князева Анна

Задачи

Создать модель машинного обучения для прогнозирования цены автомобиля:

- Выполнить разведочный анализ данных
- Выбрать параметры
- Выбрать модель и оценить её

Оценить функцию экономической амортизации автомобилей:

- Проанализировать, как стоимость автомобиля меняется со временем
- Проиллюстрировать темпы изменения

01

Анализ данных

Данные были предоставлены с сайта Kolesa.kz в формате csv-файла.

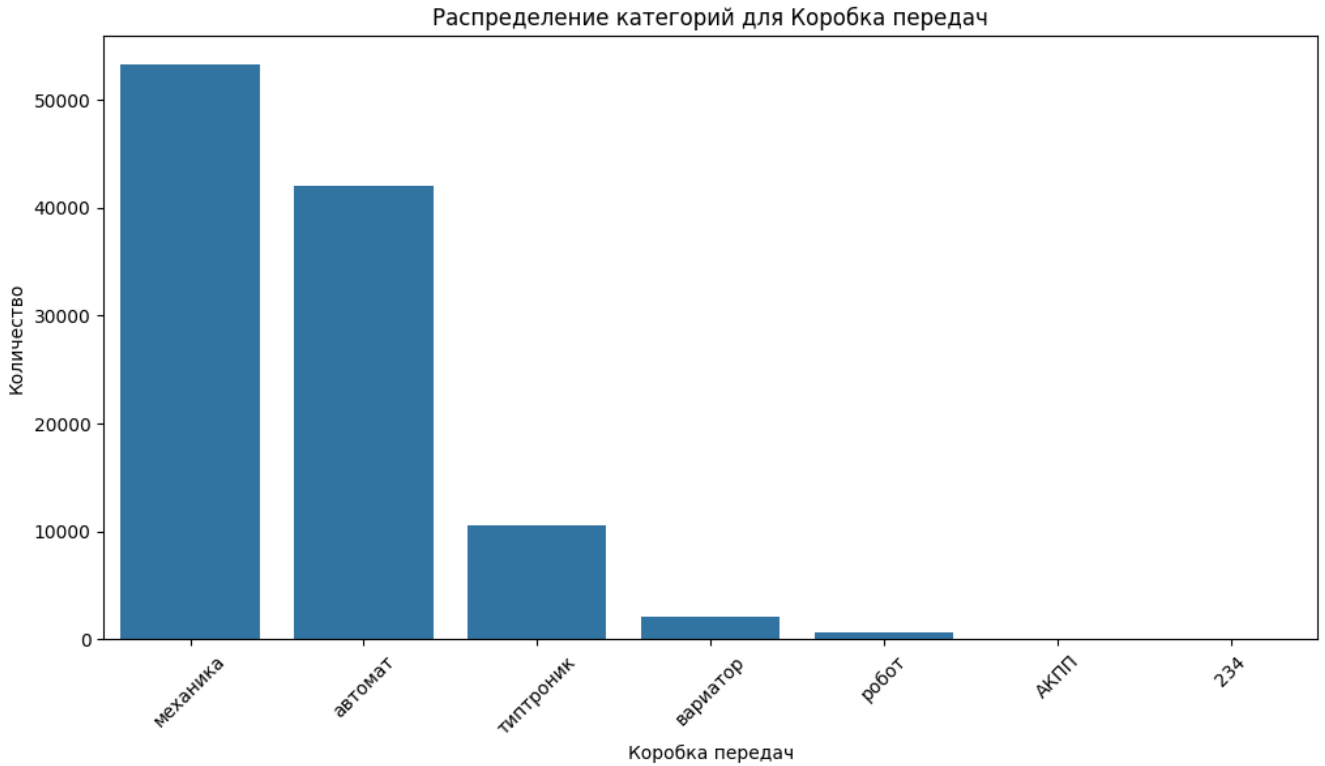
Признак «Объем двигателя» был разделён на числовой (объём двигателя) и категориальный признак (топливо).

brand	currency	description	model	price	url	year	Город	Коробка передач	Кузов	Объем двигателя, л	Привод
Toyota	₸	рейлинги, литые диски, тонировка , велюр , а...	Camry	3200000	https://kolesa.kz/a/show/95789287	2002	Алматы	автомат	седан	2.4 (бензин)	передний привод
Toyota	₸	NaN	Land Cruiser	3600000	https://kolesa.kz/a/show/95638284	1996	Актобе	механика	внедорожник	4.5 (дизель)	NaN
Chevrolet	₸	литые диски, тонировка , хрустальная оптика ,...	Cruze	3200000	https://kolesa.kz/a/show/95748872	2014	Усть-Каменогорск	автомат	седан	1.8 (бензин)	передний привод
Toyota	₸	литые диски, панорамная крыша, тонировка, люк ...	Camry	2700000	https://kolesa.kz/a/show/95781518	1998	Нур-Султан (Астана)	автомат	седан	2.2 (бензин)	передний привод

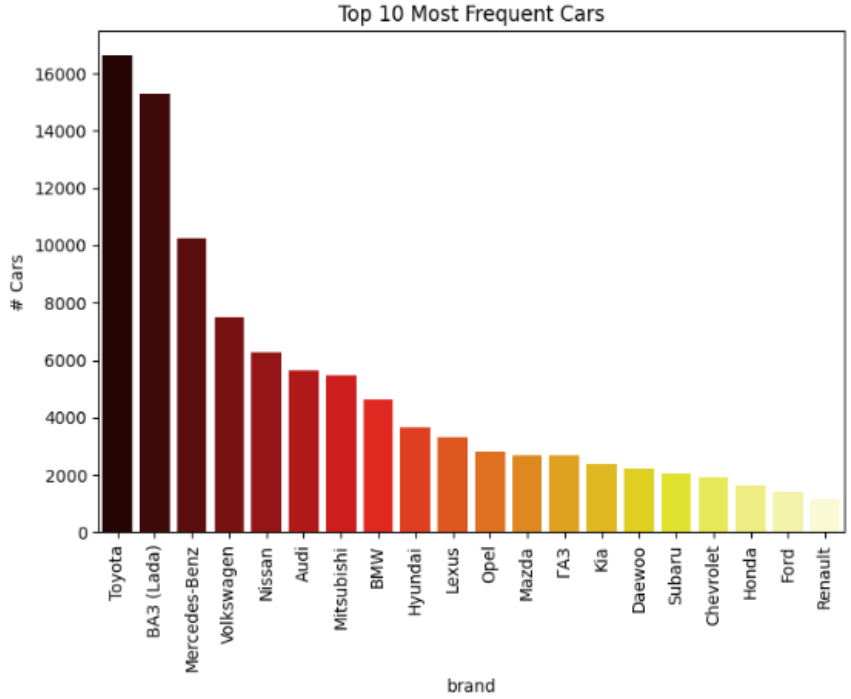
Фрагмент исходных данных

Для всех категориальных признаков построены диаграммы распределения.

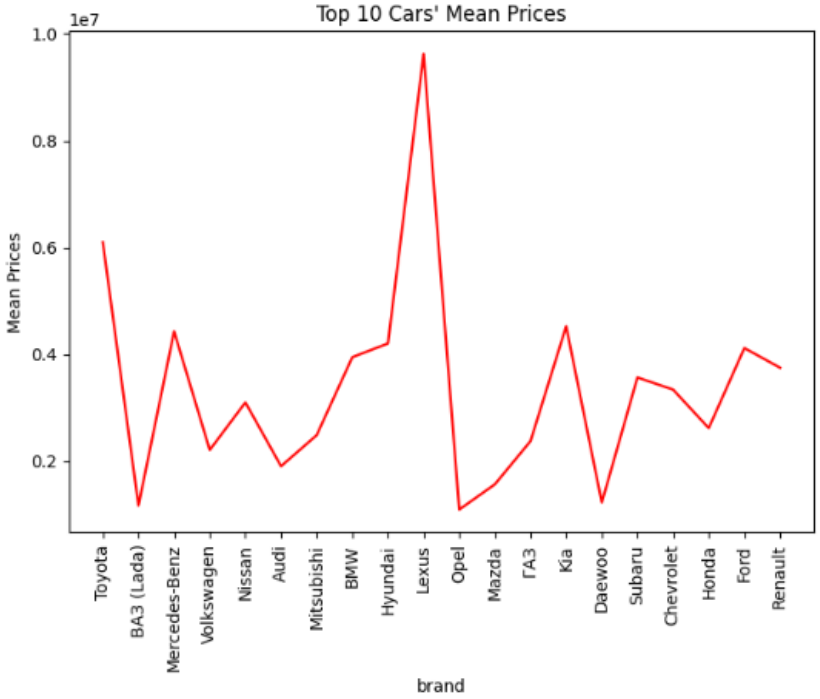
Например, для признака «Коробка передач»



Столбчатая диаграмма, отображающая 10 самых часто встречающихся марок автомобилей в наборе данных

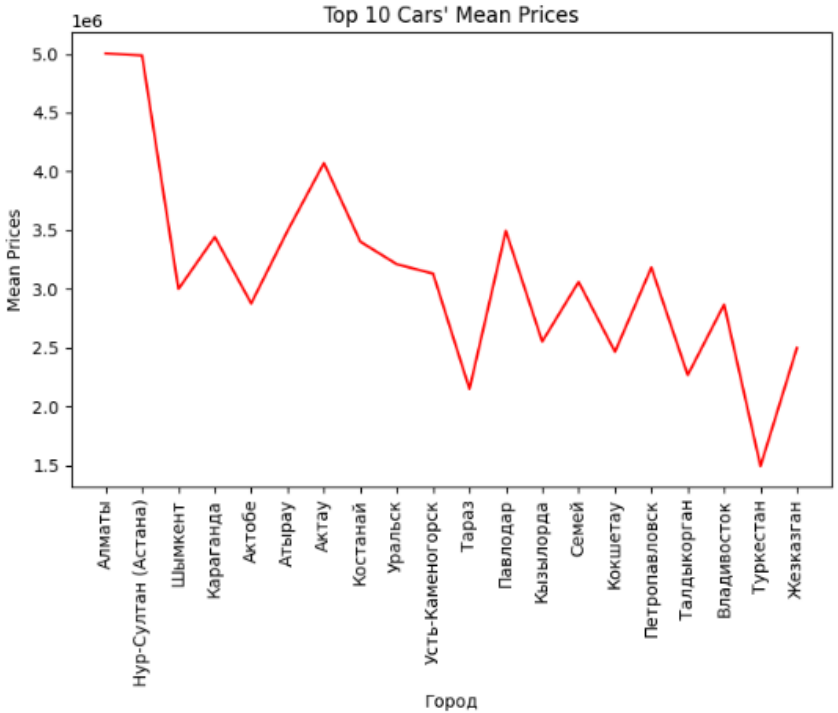


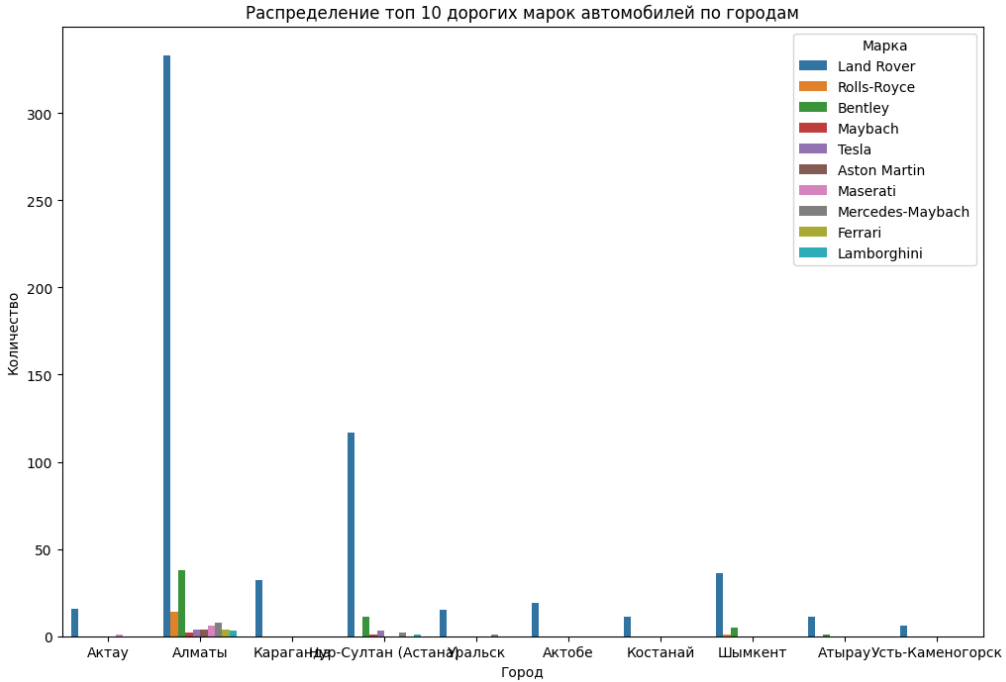
График, отображающий средние цены для каждой из 10 самых часто встречающихся марок автомобилей



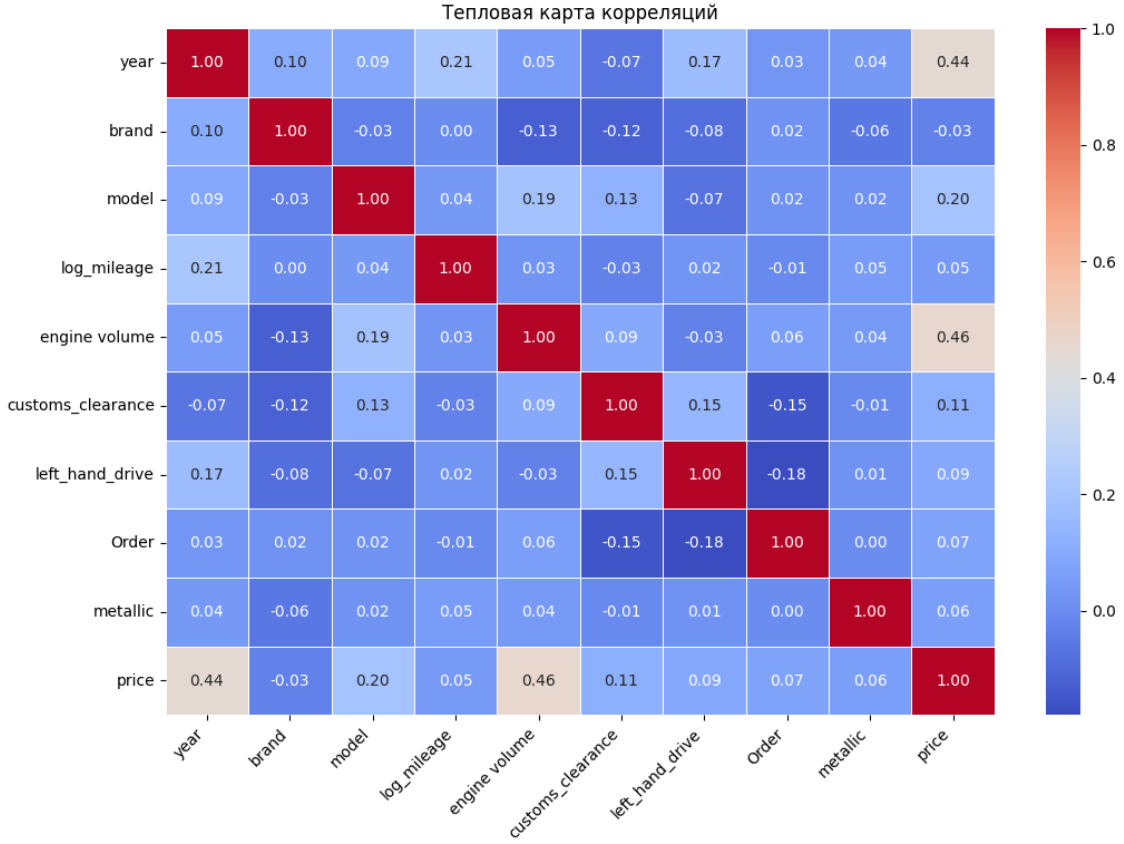
Данные графики показывают, что цены выше на автомобиле в некоторых городах, поскольку в них продаётся больше дорогих марок автомобилей.

Так что признак «Город» не был выбран для модели.



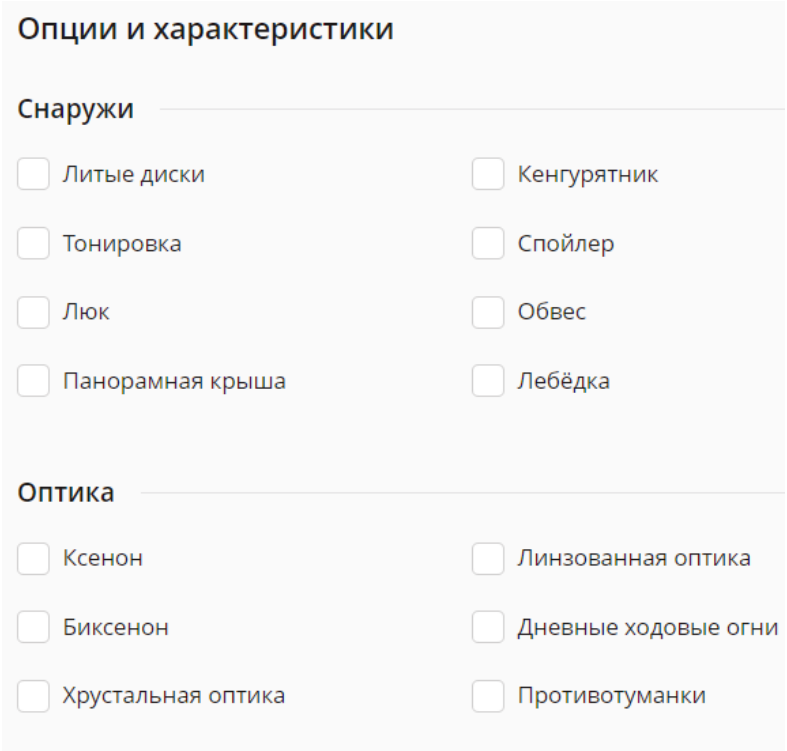


Были выбраны следующие параметры: бренд, модель, описание, год, коробка передач, кузов, объем двигателя, тип топлива, привод, растаможен, руль, цвет, пробег, наличие



Признак «description» включает в себя перечисление опций, которые пользователь выбрал на сайте и описание, которое пользователь самостоятельно ввёл.

Признак «description» был разделён на отдельные признаки по названиям опций



02

Модель

Модели

Линейная регрессия: простота и интерпретируемость, быстрота и вычислительная эффективность, возможность получить эффективную модель без погружения в подбор гиперпараметров.

Дерево решений: простота и интерпретируемость, способность к обработке нелинейных зависимостей.

Случайный лес: использует бэггинг, что позволяет уменьшить переобучение по сравнению с отдельными деревьями решений. Ансамбль случайного леса имеет более устойчивую обобщающую способность. Более устойчив к выбросам в данных.
Также для случайного леса были подобраны оптимальные гиперпараметры с помощью GridSearchCV.

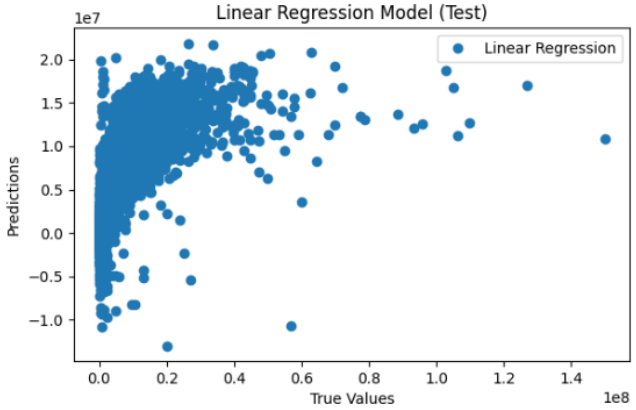
Метрики

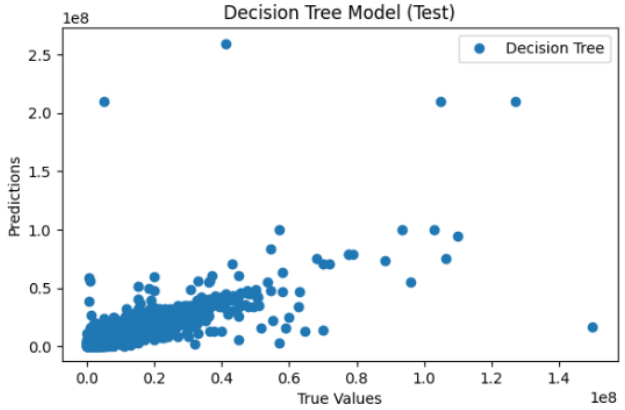
Mean Absolute Error (MAE): среднее абсолютное отклонение между прогнозами модели и истинными значениями целевой переменной. Позволяет оценить точность модели в абсолютных единицах.

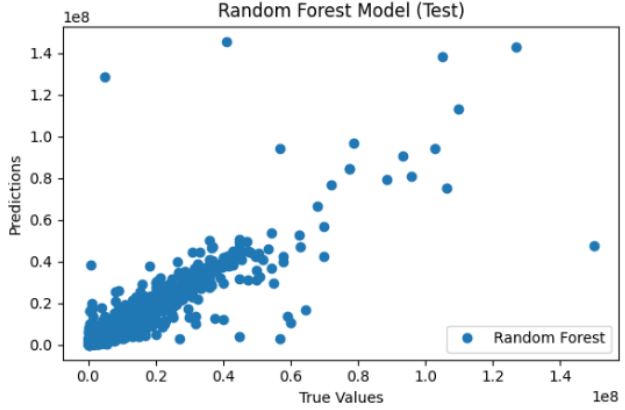
Mean Squared Error (MSE): среднеквадратичное отклонение между прогнозами модели и истинными значениями целевой переменной. Увеличивает вес больших ошибок по сравнению с MAE.

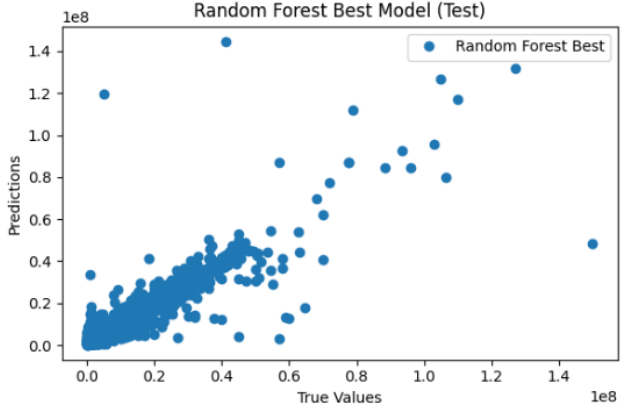
R-squared (R^2): доля объясненной дисперсии в целевой переменной. Измеряет, насколько хорошо прогнозы модели соответствуют фактическим значениям. Значение R2 близкое к 1 указывает на хорошую подгонку модели к данным, в то время как значения близкие к 0 или отрицательные указывают на плохую подгонку.

Сравнение моделей









Сравнение моделей

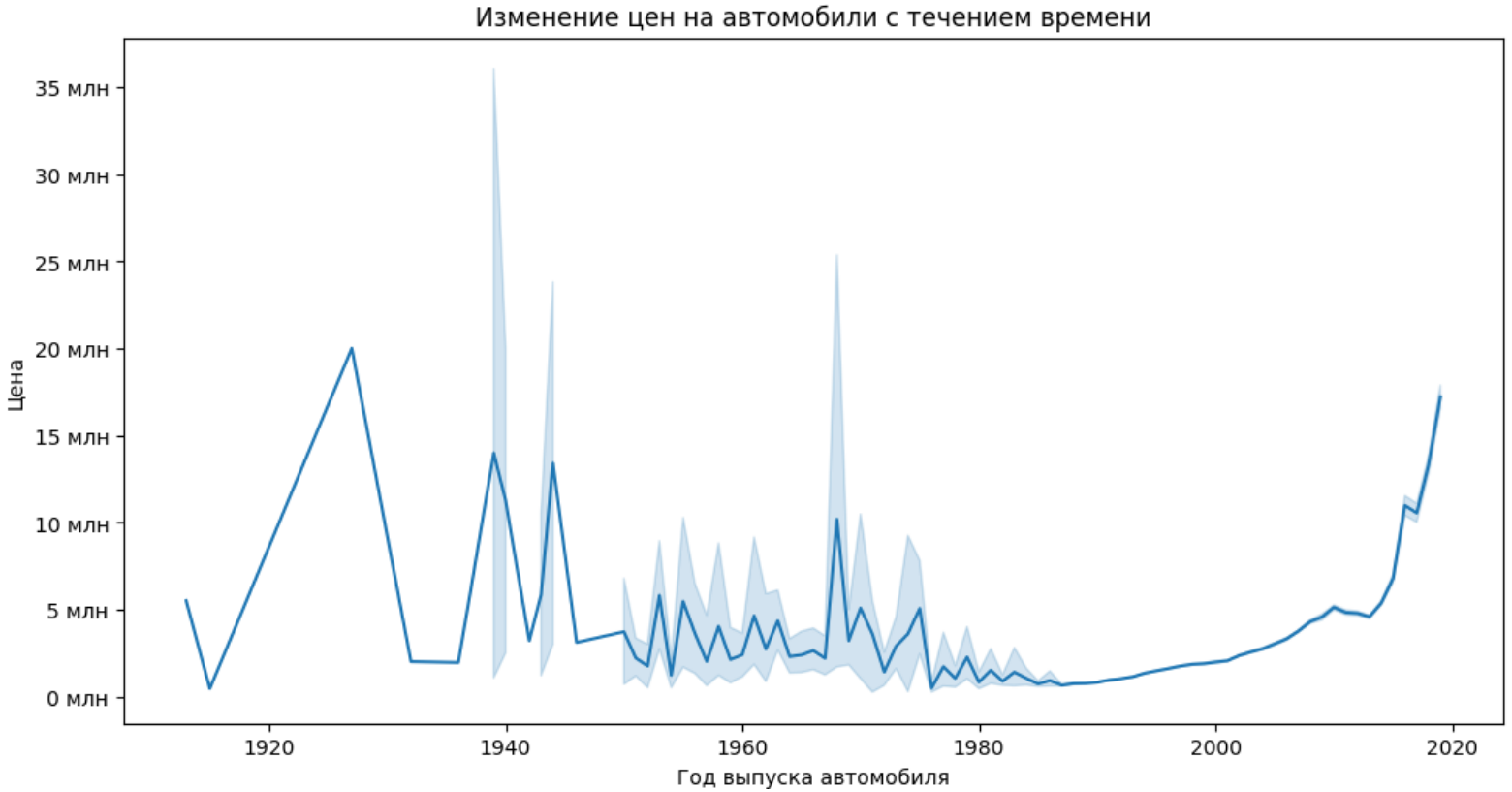
Модель \ Метрика	MAE	MSE	R2
Линейная регрессия	1,939,926	16,194,461,933,334	0.465
Дерево решений	760,418	9,801,621,581,253	0.676
Случайный лес	542,537	3,914,374,146,837	0.871
Случайный лес с параметрами	538,740	3,842,631,410,103	0.873

Выбрана модель Случайный лес с подобранными гиперпараметрами

03

Амортизация

График изменения цен на все автомобили из набора данных.

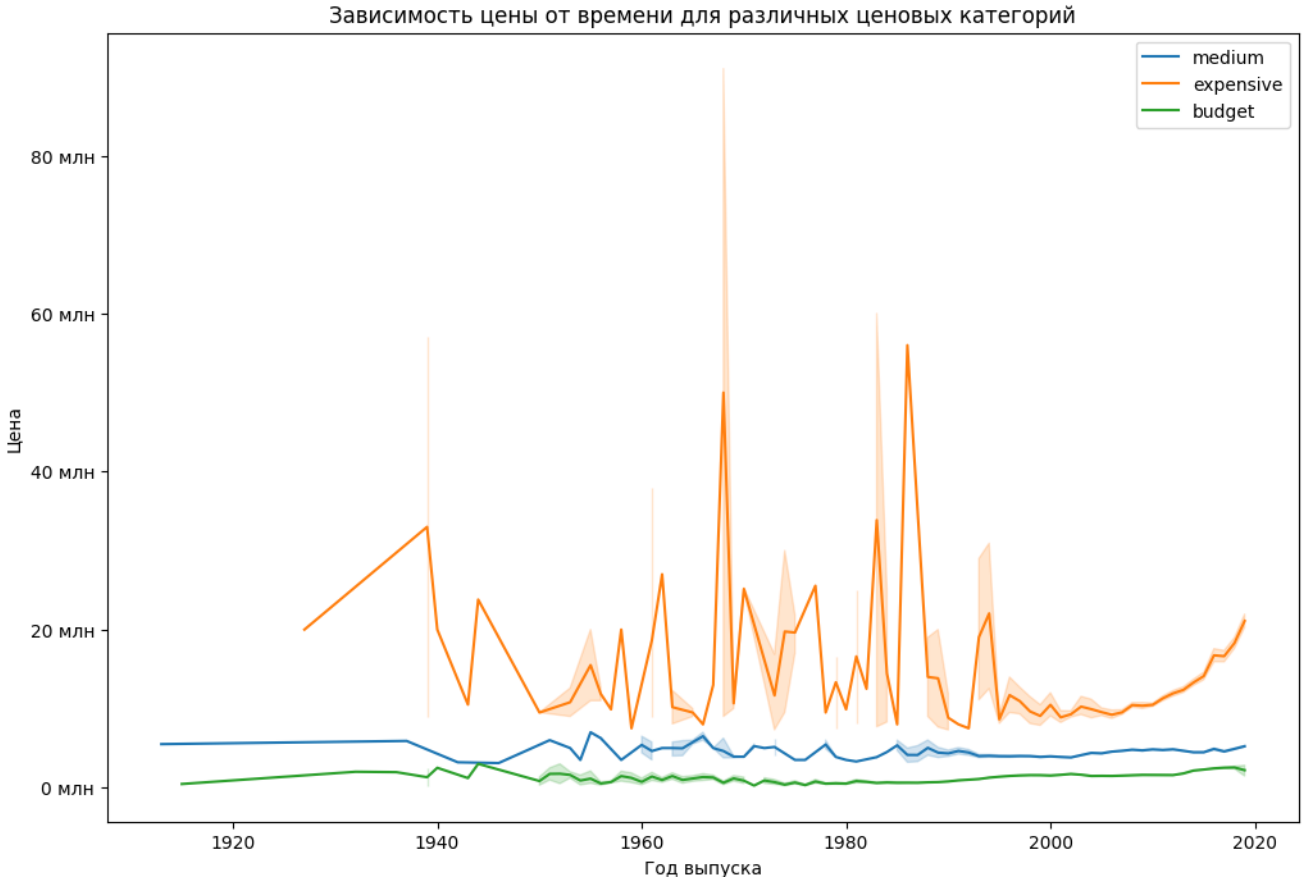


Автомобили были разделены на:

Бюджетные (до 3 млн)

Средние (до 7 млн)

Дорогие (выше 7 млн)



Бюджетные автомобили: цены на бюджетные автомобили остаются относительно стабильными с течением времени, что может указывать на низкую склонность к амортизации.

Средние автомобили: более подвержены амортизации по сравнению с бюджетными моделями. Это может быть связано с более высокой конкуренцией в этом сегменте рынка, появлением новых моделей и технологий, что приводит к более частым изменениям цен.

Дорогие автомобили: более значительные колебания цен во времени. Это может быть связано с различными факторами, такими как изменения в спросе и предложении на рынке, колебания в экономике. Дорогие автомобили также могут подвергаться более высоким процентам амортизации из-за их высокой начальной стоимости, что приводит к более значительным изменениям в цене с течением времени.