Price prediction in Second-hand car market

Presented by Team 10

Team Members

M Zaid Haris

Wenlin Zhao

MingHua Tsai

Hui Gao

Neha Shastri

Aashrith Reddy K

Contents

01 Project Background

02 Data Description & Cleaning

03 Predictive Model

04 EDA process

05 Insights and suggestions

Motivation

The U.S. used car market is a growing sector, with over 40 million cars exchanged annually, driven by increasing demand for affordable and sustainable transportation (Statista, 2023).

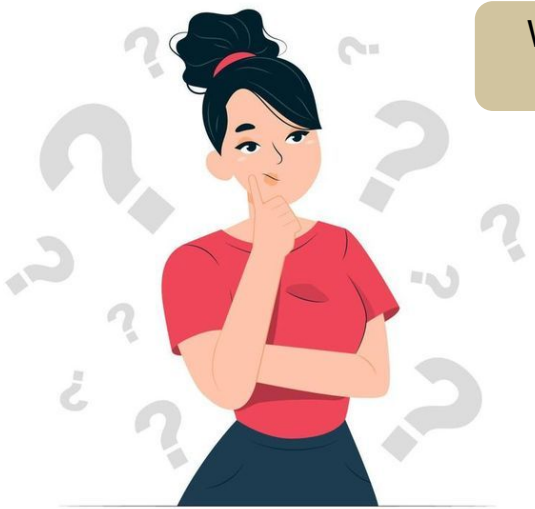
Our goal is to provide an efficient predictive model in car price for both customers and the selling platform

First of all, let's take a look into the industry

What type of car should I buy to resell at a higher price in the future?

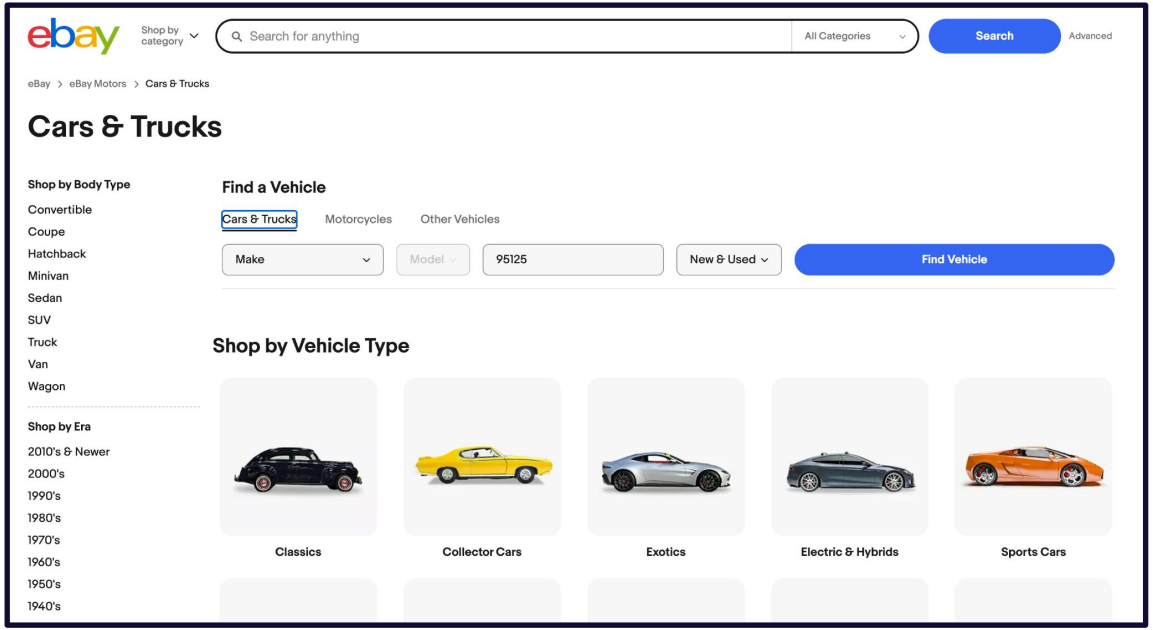
Did I get the best deal on the price of the used car I purchased?

How much can I sell my car for?



Which area can I sell my car in to get the highest price?

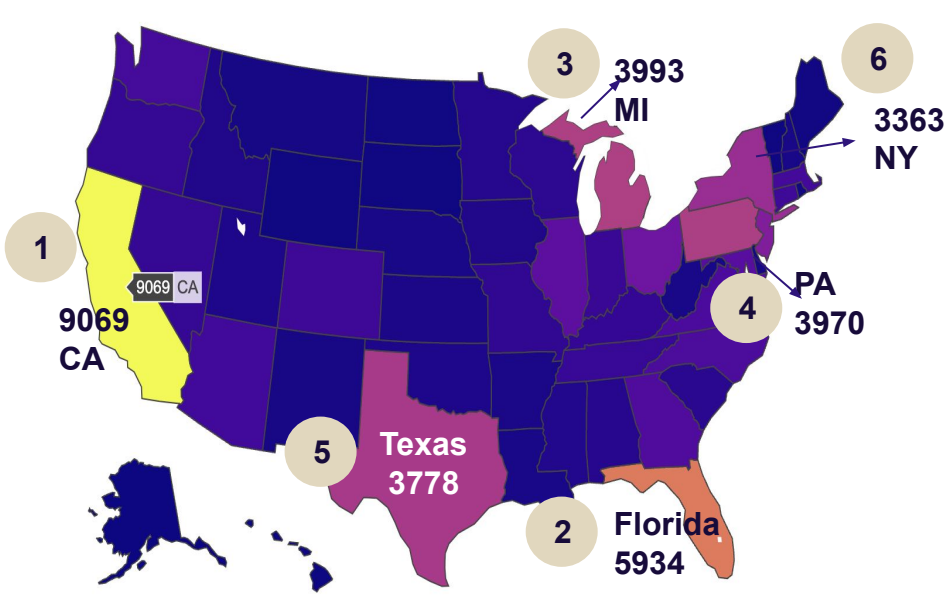
Data Resources

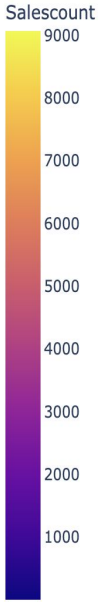


1. eBay Cars USA
2. 160k Sales Records
3. 2019 - 2020
4. Columns such as selling price, mileage, year, make

EDA & Further analysis

Let’s take a closer look at the market…



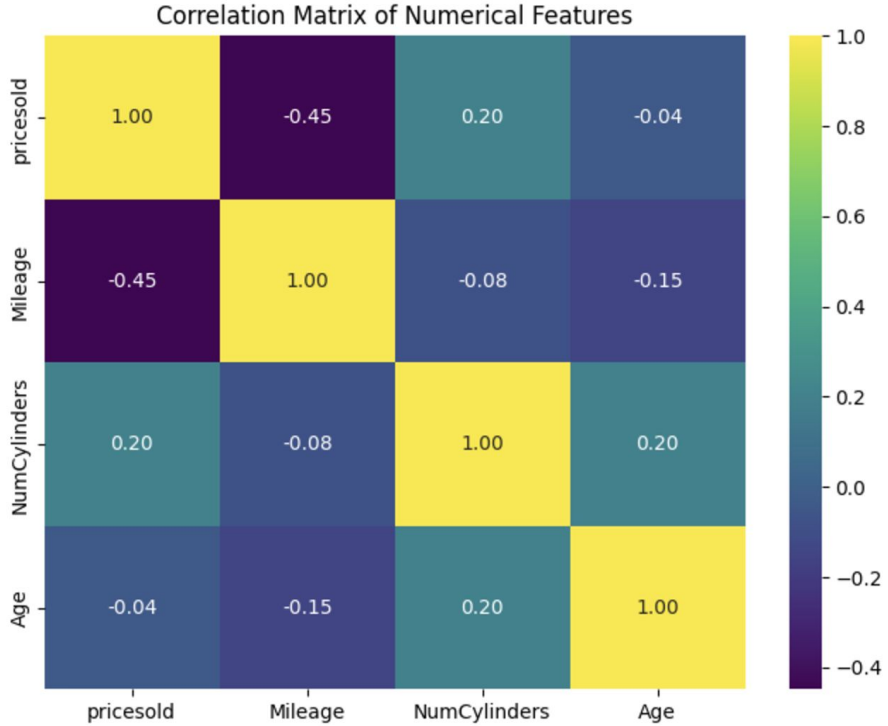


Top 3 States of sales count

1. California: The largest market for transactions in the U.S.
2. Florida: The second-largest market.
3. Michigan: The third-largest market.

California, Florida, and Michigan lead the used car market, making them key areas for focus. Pennsylvania and New York also show strong activity, while other states present opportunities for a broader strategy.

When looking into the features, we found...

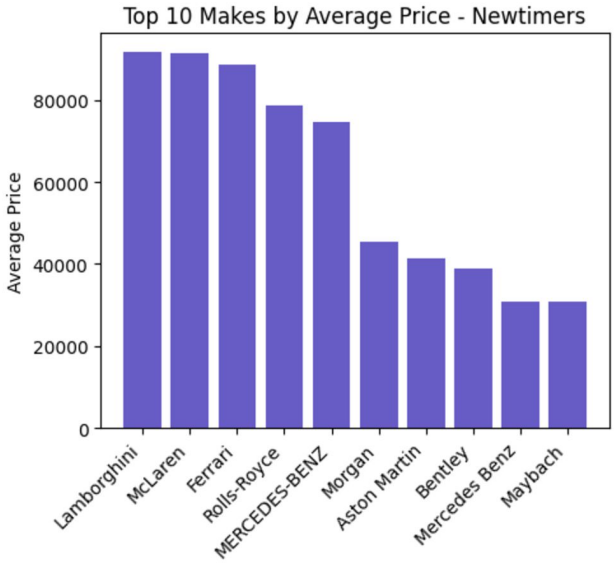


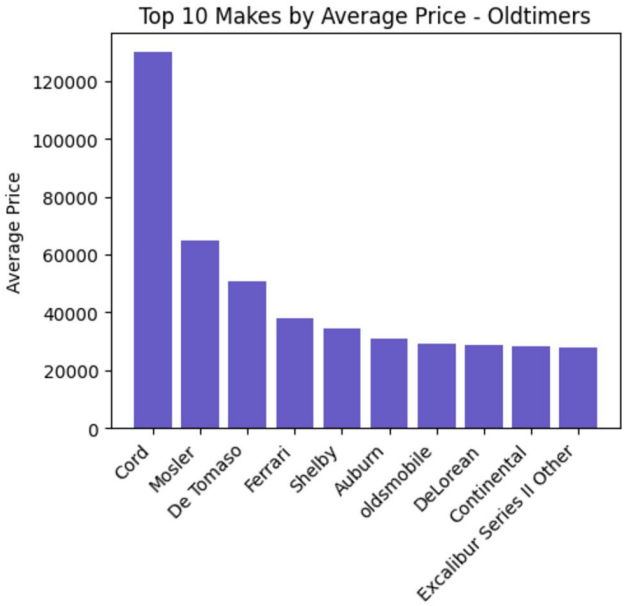
Most factors have varying degrees of correlation with price

1. Mileage having the highest impact—higher mileage leads to lower prices.

2. Age has a lower impact, contrary to common assumptions. Its influence is even less than that of NumCylinders.

Split the dataset

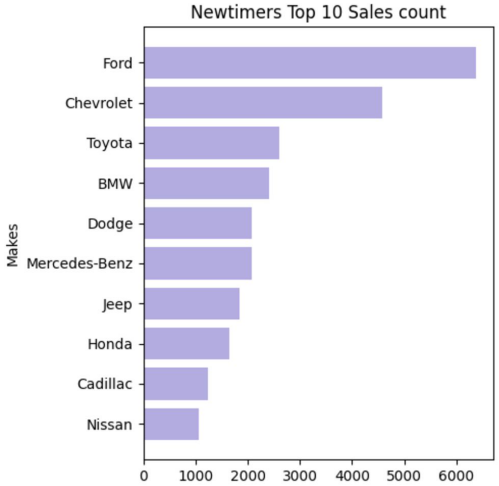


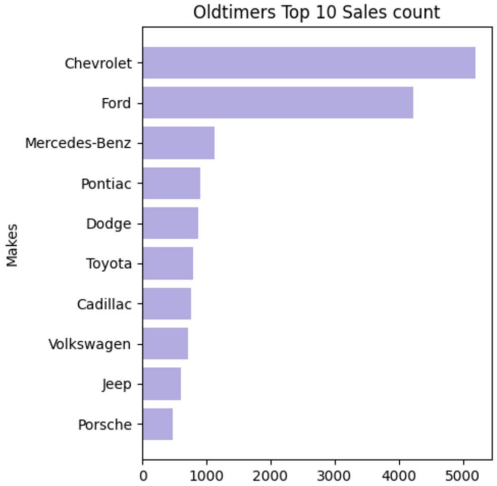


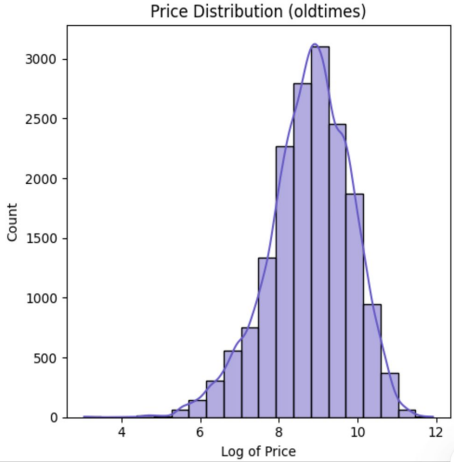
It makes sense to split the dataset into two.
1, Historical cars (age >25 years) and "normal" cars that are younger than 25 years.
2, Also, we divided into luxury cars and non luxury cars.

Cars price distribution

In terms of price, we analyzed Newtimes and Oldtimes and found difference in their price distributions.

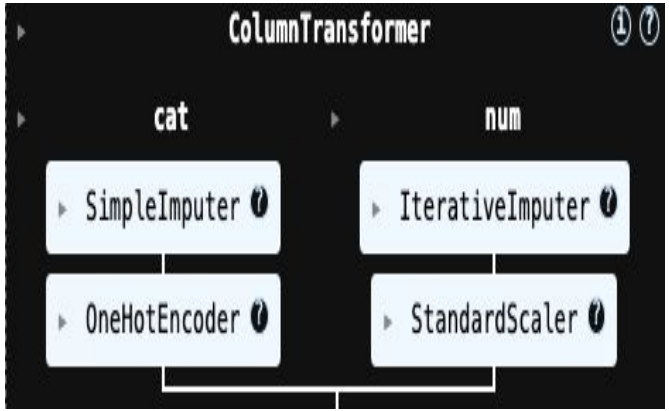






There are significant brand differences. Among the top 10 brands, Ford and Chevrolet consistently have the highest transaction volumes, while Toyota performs better in Newtime, and Mercedes-Benz sees higher transaction volumes in Oldtimes

Data Preprocessing



- **Splits dataset into 2 groups**

Categorical

- **SimpleImputer: Handle missing value by filling them(frequent value/ constant)**
- **OneHotEncoder: Converts categorical variables into binary columns**

Numerical

- **InteriativeImputer: Predicts and fills missing numerical values using other features**
- **StandardScaler: Scales numerical features to have a mean of 0 and standard deviation of 1**

```
preprocessing.get_feature_names_out()
#get number of features
len(preprocessing.get_feature_names_out())
```

```
16149
```

Handles high-dimensional data effectively
Need dimensional reduction / feature selection model — improve performance

Model Process

Model Selection

- Linear Regre,Ridge Regre, Decision Tree
- Advanced algorithms: XGBoost, Catboost

Hyperparameter Optimization

- Improve accuracy: Halving Random Search

Feature Importance Analysis

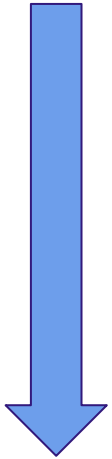
- Based on model results, we just find what features will influence the car price

Model Training and Prediction

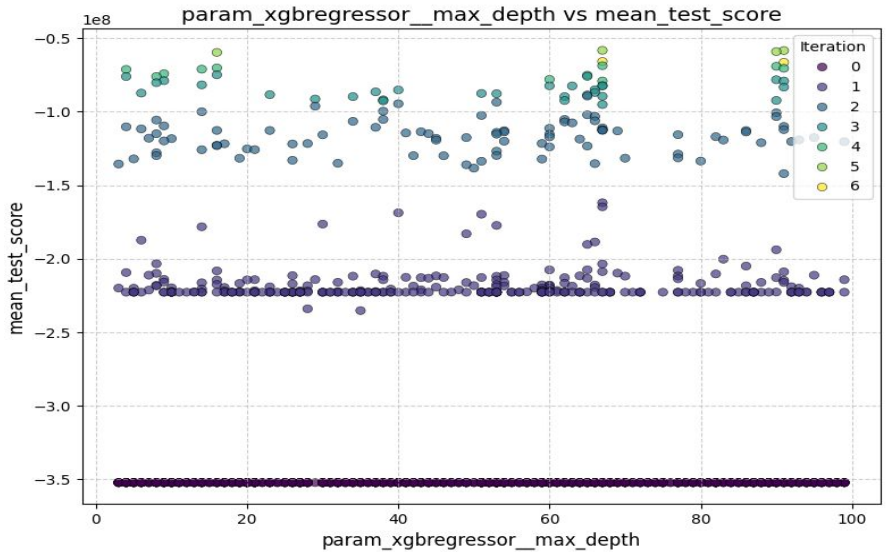
- Final models were trained using optimized hyperparameters
- Predictions were made based on numerical and categorical features

Model Selection

Model Choose	RMSE Value
Avg Linear Regression Cross-Validation	8799
Avg Decision Tree Regression Cross-Validation	8625
Avg Ridge Regression Cross-Validation	8350
Average Gradient Boosting Regression Cross-Validation	7863
Average LightGBM Regression Cross-Validation	6948
Average **XGBoost** Regression Cross-Validation	6833
Average **CatBoost** Regression Cross-Validation	6822



Hyperparameter Tuning - Halving Random Search



Catboostregressor hyper-parameters:

- depth: 13
- iterations: 482
- l2_leaf_reg: 1.8011484638467514
- learning_rate: 0.09948166560605276
- random_strength: 5.5759237615373625

XGbregressor hyper-parameters:

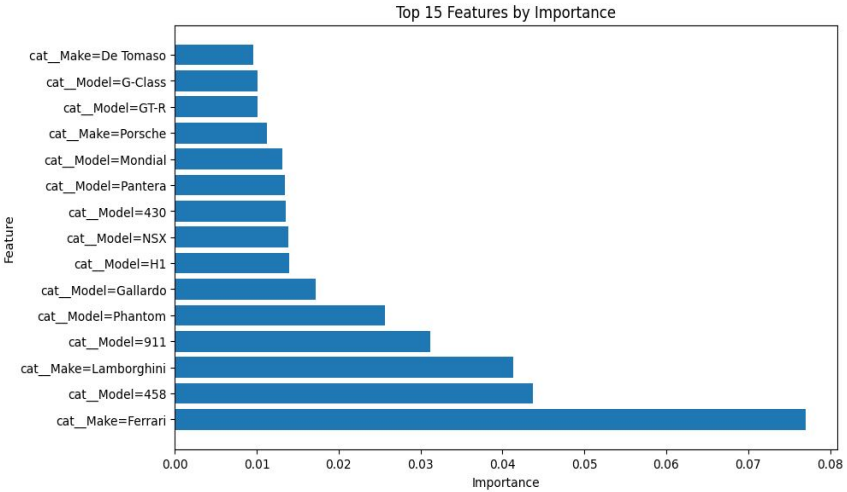
- colsample_bytree: 0.603574744640959,
- gamma: 1.1042639148150446,
- learning_rate: 0.01350369669931007,
- max_depth: 67,
- min_child_weight: 5,
- n_estimators: 377,
- reg_alpha: 1,
- reg_lambda: 8,
- subsample: 0.569487246386216

Catboost RMSE = 6395

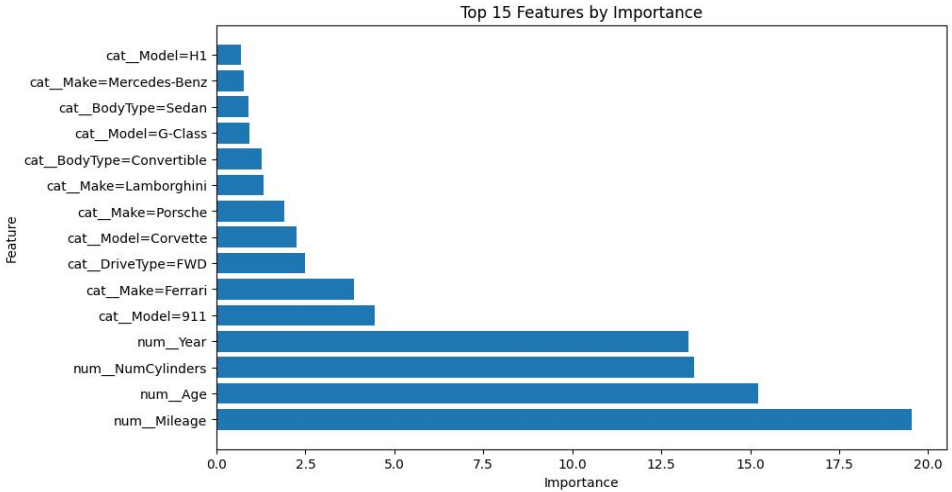
XGBoost RMSE = 6497

Null Model RMSE = 11866.67

Top 15 Features

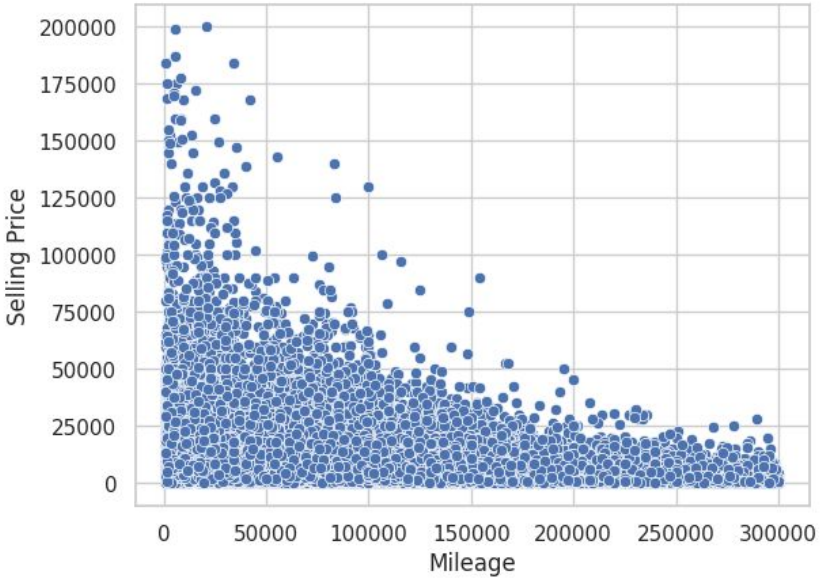


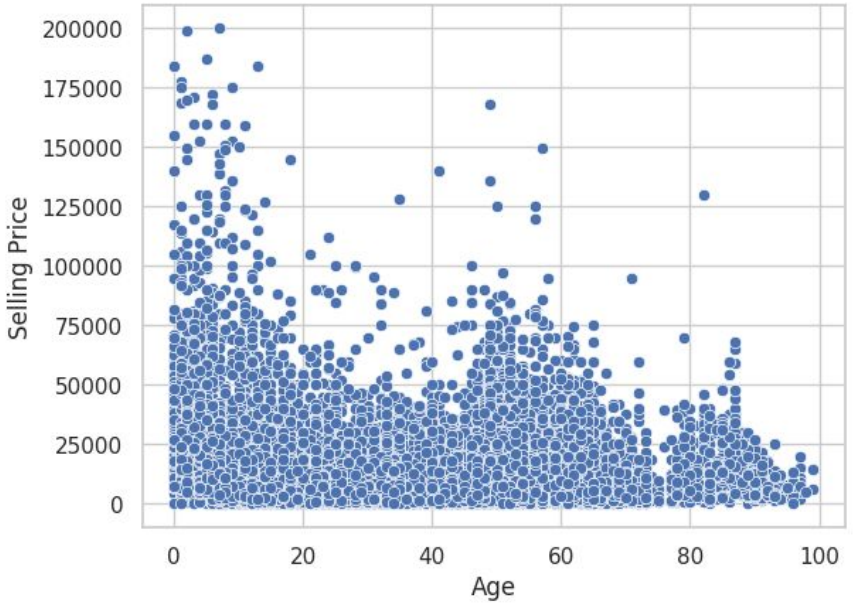
CatBoost



XGBoost

How does vehicle condition matter?





How can we bring value to customers?

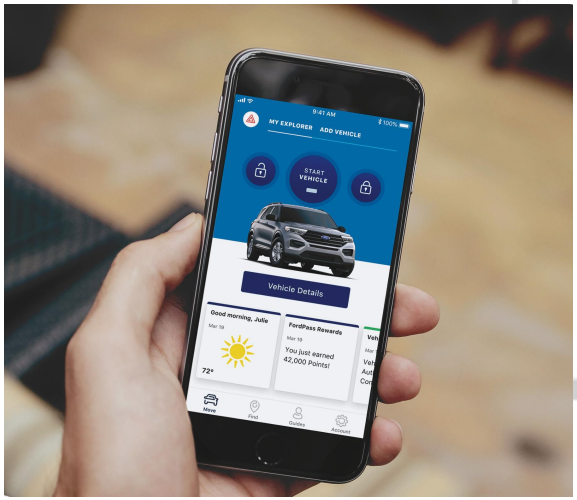
Did you know?

The US automotive industry was valued at $1,514.8 billion!
The US used car market was valued at $201.2 billion!

Companies willing to buy your car - Carvana, eBay, CarGurus
‘Should I buy the car today?’

Recommendations & Future Steps

A New Website Platform to Solve Problems!



SmartCarDealers

Your Best Second-Hand Car Selling Partner

- **Car Price Prediction**
- **Forum And Data report**
- **Car Market**

Secure, Transparent, and Efficient !

Challenges

Vehicle-specific details:
VIN number, accident history, and maintenance records.

Seller behavior:

1. Maximize profits by setting higher prices vs. A quick sale
2. Large dealerships like CarMax: Provide upfront price based on basic vehicle details.
3. Smaller dealerships: Employ a more detailed process

Thanks!

Q&A