Car price: EDA & Prediction

1. Problem statement

The lack of comprehensive data and price uncertainty for new and used cars have become significant concerns in the current automotive market. Several factors, including rapid technological advancements, fluctuating market conditions, and changing consumer preferences, contribute to this issue. Additionally, inconsistencies in car valuation methods and limited access to historical pricing data further exacerbate the problem.

As a result, determining the appropriate pricing for new and used cars has become increasingly challenging. Sellers face the risk of undervaluing their vehicles, leading to potential revenue loss, while buyers may be deterred by overpriced cars or feel dissatisfied after purchase. This situation calls for improved data collection, analysis, and sharing to establish fair and accurate car pricing, ultimately benefiting all stakeholders in the market.

2. Solution Approach

Three-step solution approach was adopted:

a. *Data Collection through Web Scraping*: Leverage web scraping tools to collect current market data about price sales and car features from automotive marketplaces.
b. *Data Analysis*: This step involves identifying patterns, correlations, and trends that can impact car prices. Additionally, remove any inconsistencies or outliers, and handle missing values to ensure the dataset is reliable and accurate for developing a predictive model.
c. *Machine Learning Model for Car Price Prediction*: Pipeline with machine learning model and sample web app.

3. Stakeholder and benefits

This project was developed for a local car agency located in Argentine. The app provides accurate price of reference of new and used car, which implies faster sales for the company through increasing the likelihood of attracting potential buyers. The insights provided by the model enable sellers to make informed decisions about vehicle offerings, promotions, or trade-in valuations.

3. Data

Price sales data of new (0 km) and used cars and their features were obtained from Mercado Libre (The most significant marketplace in South America). A specific web scraping algorithm was developed to get the price and features of the published car (web scraping script can be accessed here). The web scraping script works based on the following steps.

a. The user introduces the product to search and geographical location and country.

b. The script searches in the marketplace browser, getting the results on multiple pages.

c. The script accesses the results links to get each item's price, car features, and photos through HTML parsing.

d. The features are organized into a dictionary.

e. Data is transformed into a CSV file.

In this sample, the data obtained is geographically limited to the city of Buenos Aires (the Capital of Argentina). However, the methodology presented in this work can be applied to different cities or countries.

4. Tools

Web scraping, Python, Streamlit, BeatifulSoup, Data Visualization

5. Results.

6.1. Exploratory Data analysis (EDA)

Multiple data cleaning techniques were applied to the raw data, which presented severe issues related to the data format, value inconsistency, price sales expressed into different coin exchange or format inputs, missing values, or unavailable values for other categories. The complete data cleaning script of raw data can be accessed here. The Argentinian currency (Peso) price in raw data was transformed into American dollars, providing a better price reference. A conversion of 1 U$S = 382 ARS was adopted. The new and used car sale price distribution is shown in Fig 1.

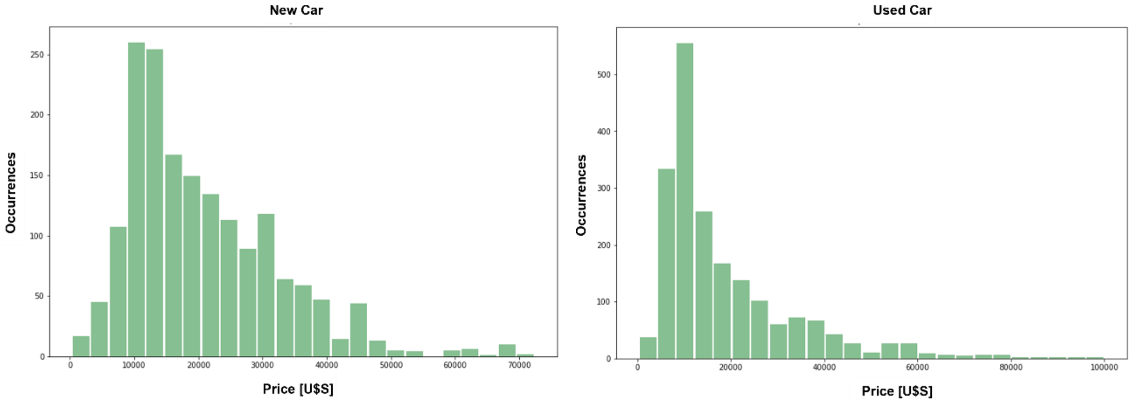

Fig 1. Sale Price distribution of new and used cars.

As shown in Fig 1, the highest prices of occurrences were found for a price range between 8000 and 12000 U$S and 5000 and 8000 U$S for new and used cars, respectively. Furthermore, both data exhibit longtail, indicating a wide divergence in car prices. However, a longtail is more pronounced in a used car than in a new one. This observation agrees with the assumption that new cars are more expensive than used cars. Under this perspective, used cars with prices higher than 70000 U$S (highest new car price) could be considered an exception, or used cars with particular features, for example, limited edition or luxurious cars. Commonly, the version car has a strong influence on its value. Fig 2 and Fig 3 exhibits the top 30 car companies and models for the new and used car, respectively.

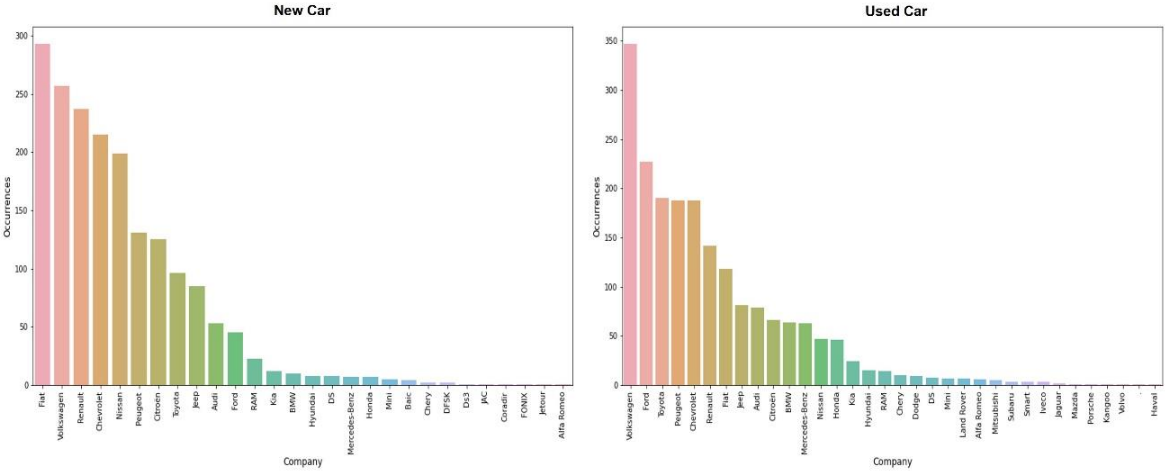

Fig 2. Top 30 companies of new and used cars.

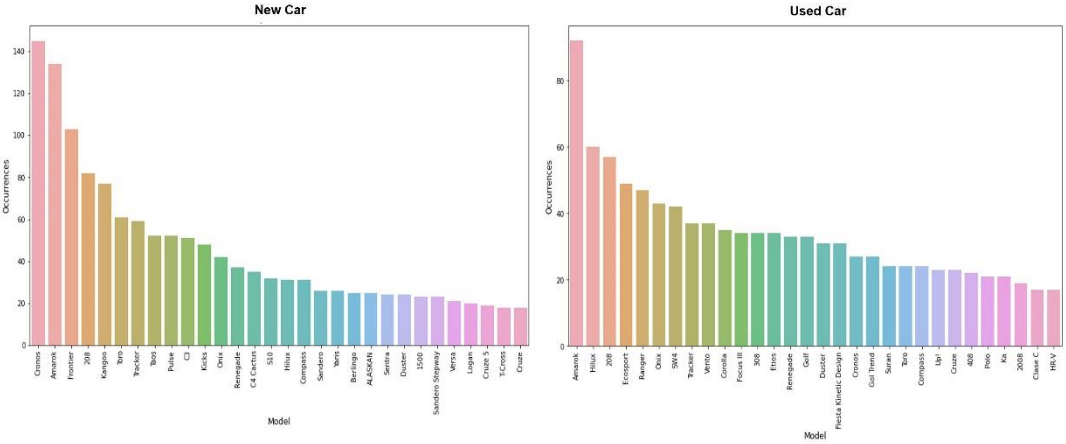

Fig 3. Top 30 models for new and used car.

According to the results, Volkswagen has more car posts in the used car category, followed by Ford, Toyota, Peugeot, and Chevrolet. On the other hand, Fiat is the top company for new cars, followed by Volkswagen, Renault, Chevrolet, and Nissan. Data related to new and used Cars produced by luxury companies are scarce (e.g., Alfa Romeo, Volvo, and Jaguar). Regarding used cars, Amarok, Hilux, and 208 are the top models published. Meanwhile, Cronos, Amarok, and Frontier are the top three models in the 0 km category. Notably, “1.3 Gse Drive Pack Conectivity” and “2.8 Srx 177 cv 4x4 7as At” are the two cars more published in the marketplace for the new and used car, respectively (Fig 4).

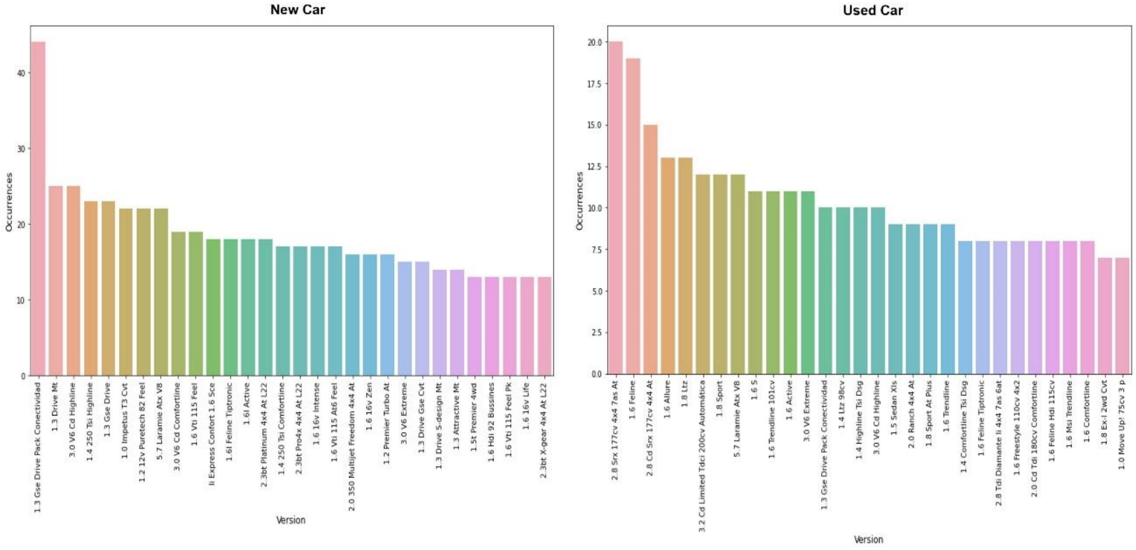

Fig 4. Top 30 version for new and used car.

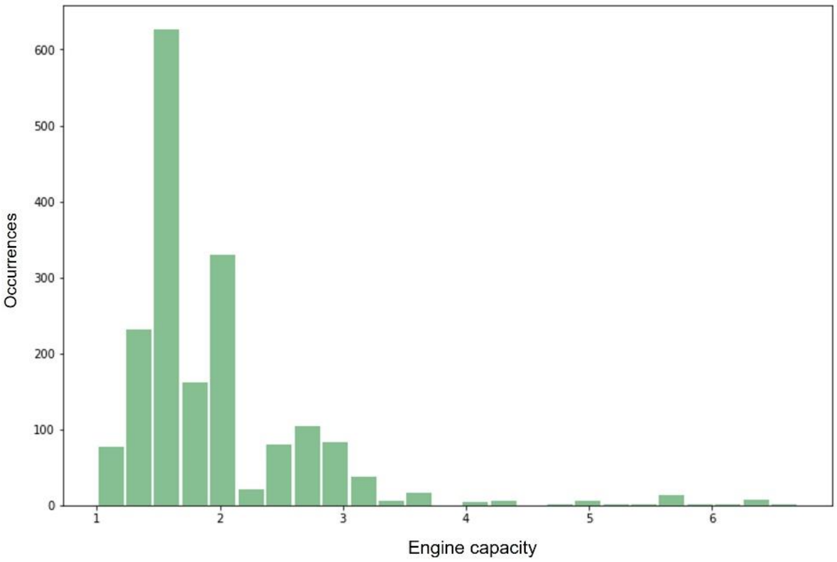

Fig 5. Engine capacity of used car.

Fig 5 shows the engine capacity distribution in a used car, exhibiting a broad range between 1.0 and 6.5. The most common engine capacity range is from 1.3 to 3.0. Used vehicles with an engine capacity higher than 3.5 are scarce. Items with automatic transmission represent 42% and 54% of used and 0 km cars, respectively. This observation indicates that published new vehicles tend to replace manual transmissions with automatic transmissions. The complete EDA (Exploratory Data Analysis) can be accessed here.

6.2. Prediction and MVP

Once the data was cleaned and performed the EDA, a linear regression model was developed using a pipeline. This pipeline was deployed using FastAPI. Also, a Minimum Viable Product (MVP) was designed using the Streamlit framework. In this app, the user introduces the features of the car, and the price of the user or new vehicle is displayed (Fig 6). The web app instructions can be observed here.

Fig 6. Caption of MVP app for car prediction.

6. Conclusions

a. Mercado Libre marketplace is a valuable and updated data source of the car sales market based on their features.
b. Multiple data cleaning techniques and EDA were applied to find the different trends in raw data.
c. The highest prices of occurrences were found for a price range between 8000 and 12000 U$S and 5000 and 8000 U$S for new and used cars, respectively.
d. Volkswagen has more car posts in the used car category, followed by Ford, Toyota, Peugeot, and Chevrolet. Fiat is the top company for new cars, followed by Volkswagen, Renault, Chevrolet, and Nissan.
e. Data related to new and used Cars produced by luxury companies are scarce.
f. Items with automatic transmission represent 42% and 54% of used and 0 km cars, respectively.
g. Price prediction using a pipeline with linear modeling was developed. Satisfactory prediction metrics were reached.

8. Further works

a. Analysis of different algorithms to test the prediction accuracy.
b. Thorough EDA between the different car features based on stakeholders' needs.
c. Study the location variable (city, countries) effect on new and used car prices.
d. Incorporation of extra functionalities to the MVP, for example, comparative charts.