Used Car Price Range Prediction

Anthony Kwok


USED
CARS

Story Behind

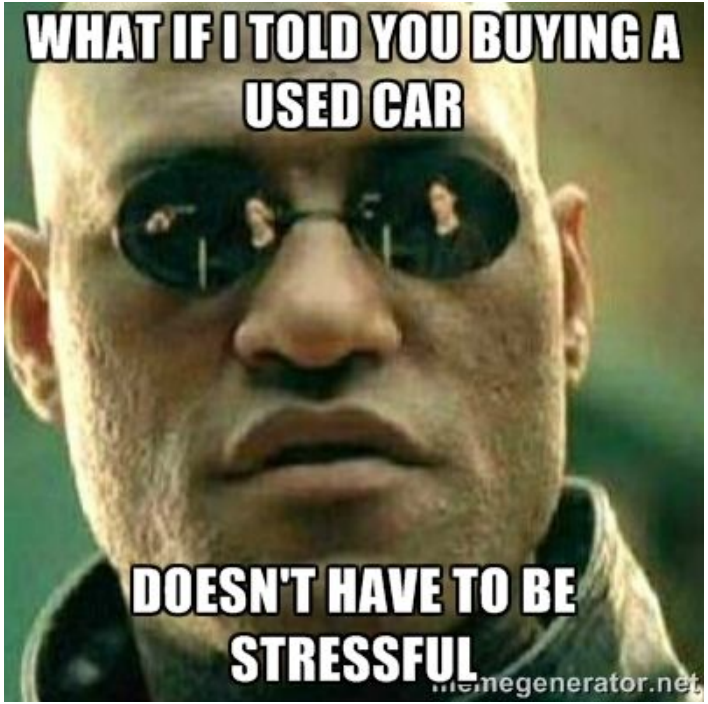
- My friend, Jason, wanted to buy a car
- Went to dealers for used car listing
- Spent 4 months with no feedbacks

- Found one on craigslist
- Within budget & requirements

Problem Statement

What we need it is

Accurate & Transparent Pricing



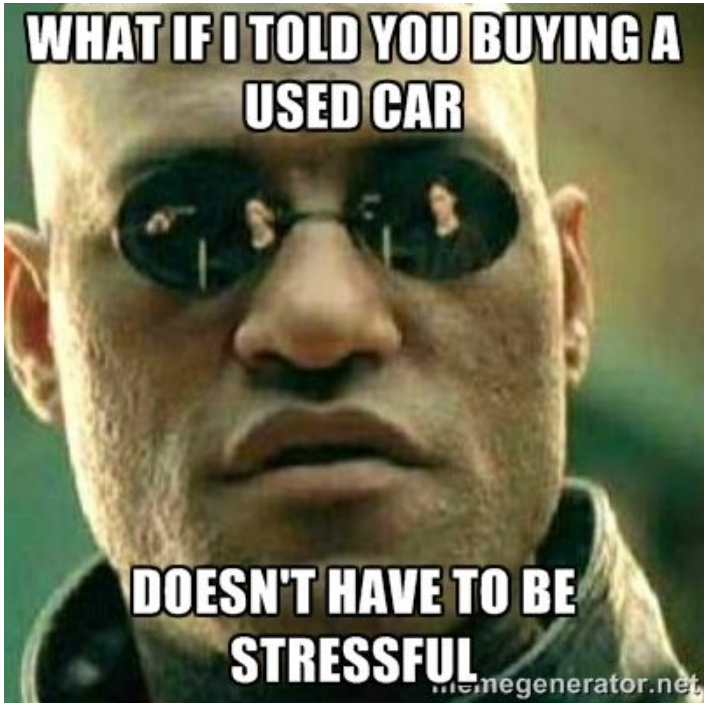
Problem Statement

Buyer: Difficult to assess if the listed price is reasonable

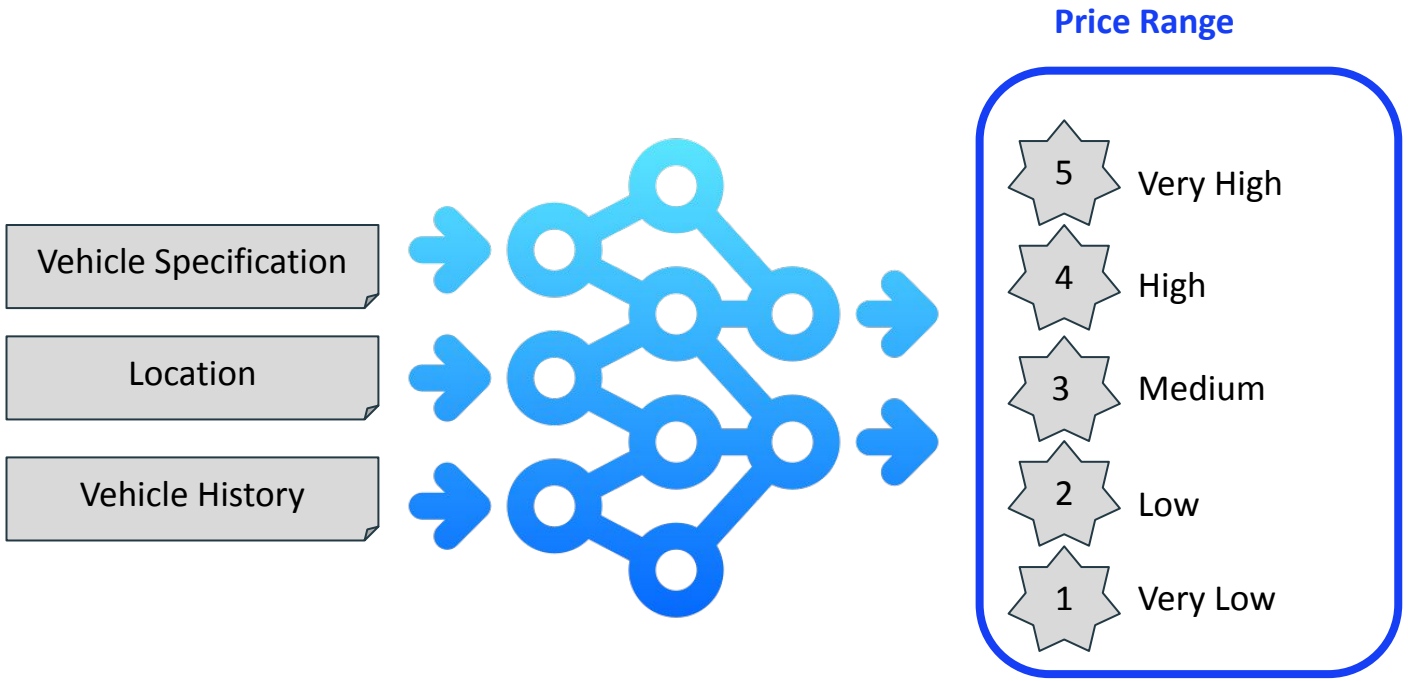
⇒ price uncertainty and the potential for overpayment.

Seller: Struggle to set competitive (higher) and attractive (lower) listing prices

⇒ unsuccessful sales or missed opportunity



Data Science Solution



Potential Impact

According to ***Canadadrives***, it takes up to 4 weeks to sell a used car.

- With our prediction model, we can provide a reasonable valuation of the used car to both buyers and sellers. ⇒ **Increase Market Efficiency** ⇒ Reduce 10% of operational cost.

Assume being a part of Data Team in a Car Listing Online Platform,

- This ***price range prediction service*** could be a premium feature that requires extra transaction fee⇒ Extra 10% of gross profit

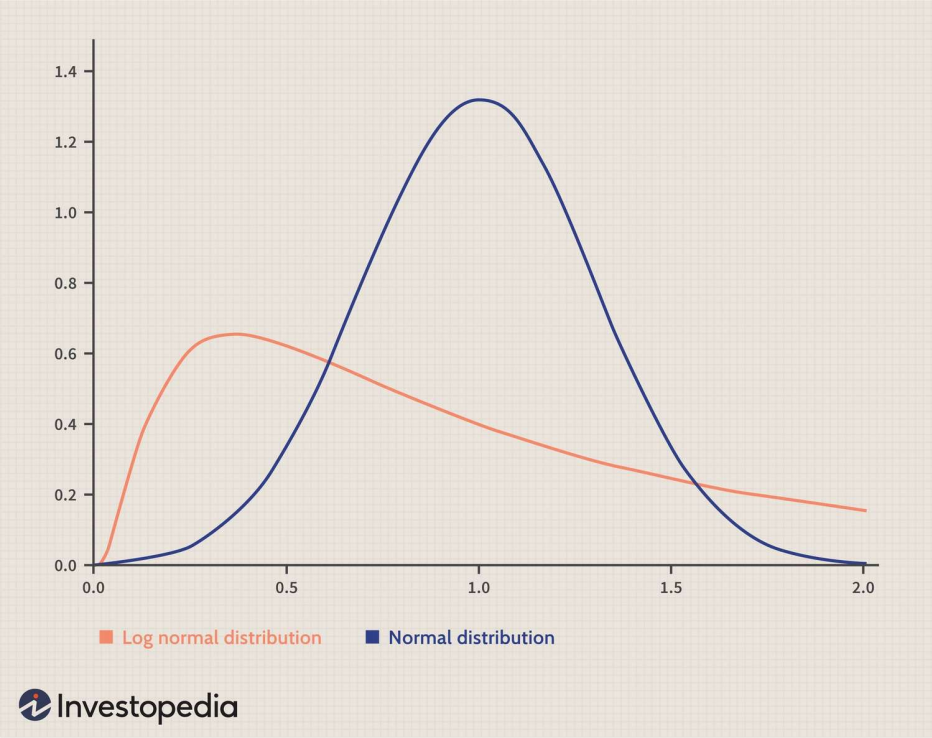
- In Total, bring up extra 20% of gross profit (reducing 10% of cost + increase 10% of gross profit.)

Dataset

- From Kaggle
- Two files
  - CA
  - US
- Total ≈7.5M rows
- After cleaning,
  - 13 numerical columns
  - 12 categorical columns
  - 1 target variable

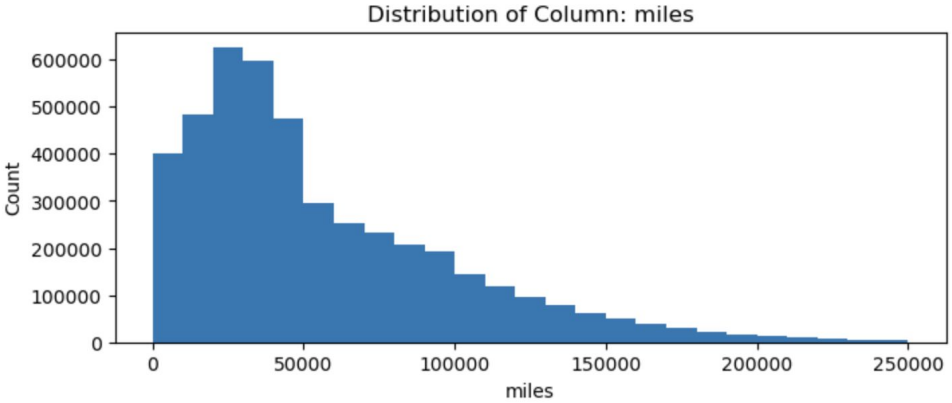
EDA Findings (1)

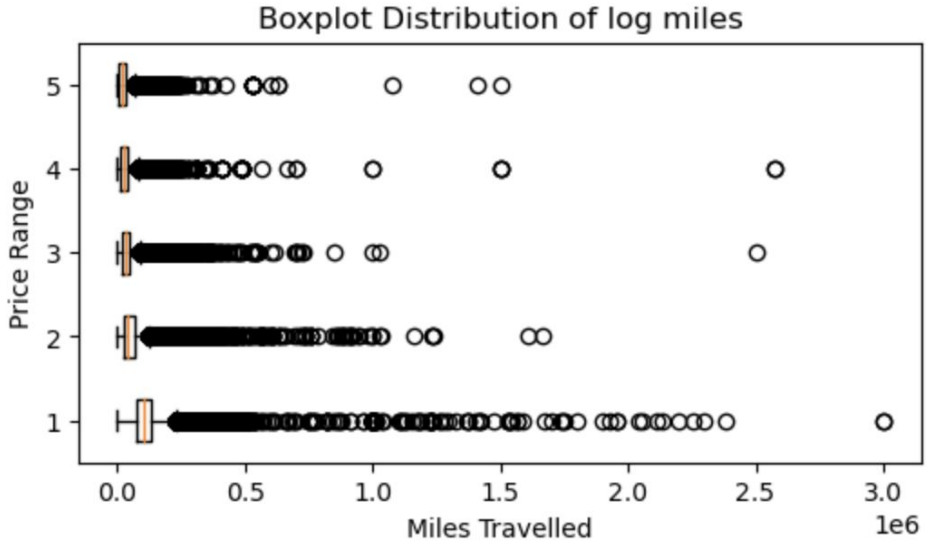
- Normal Distribution
- Log-Normal Distribution



EDA Findings (1)

- Column: “miles”
- Looks like log-normal distribution
- Try to apply log on x-axis

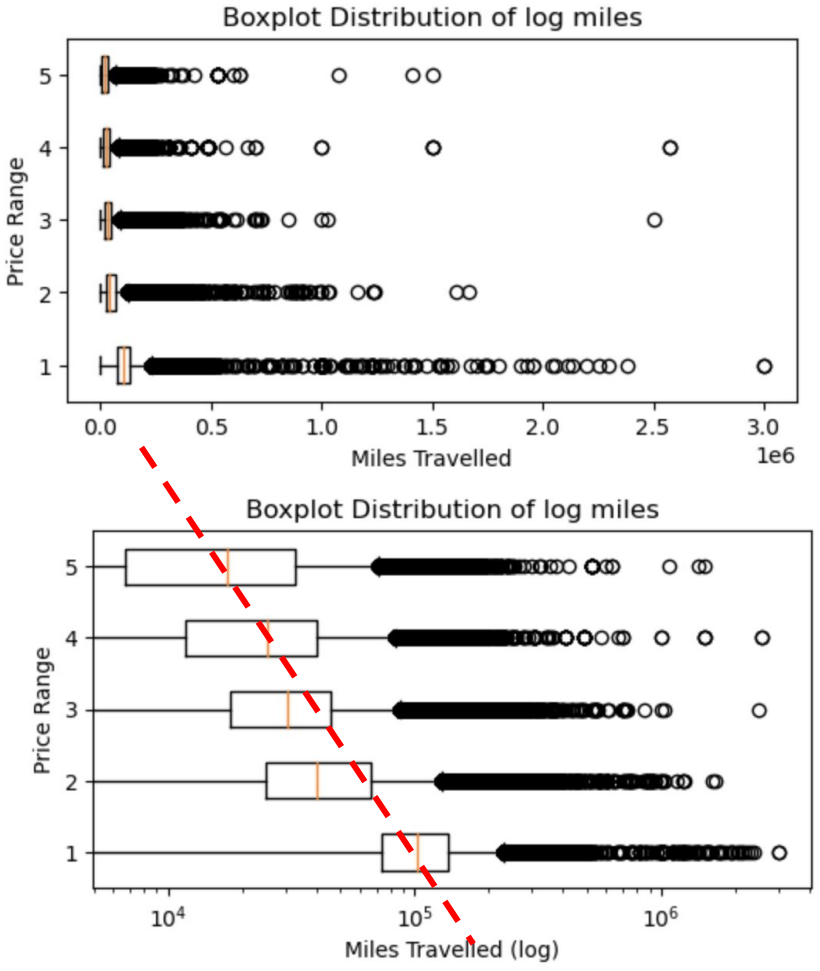




EDA Findings (1)

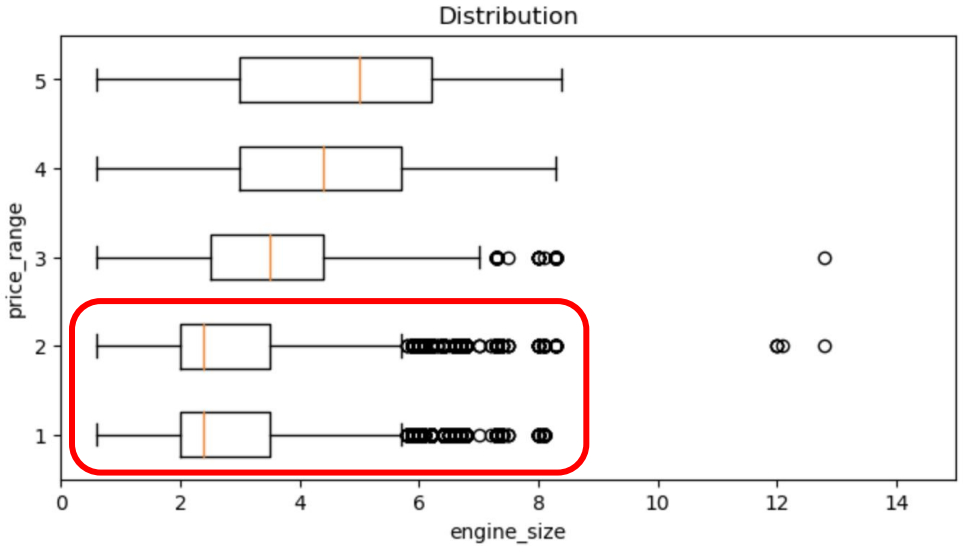
- “Price range” seems to be negatively correlated to the “miles travelled” (red line)

- Apply ANOVA test for further investigation



EDA Findings (2)

- For price range group 1 & 2, they are very similar. (red box)
- For group 3 & 4 & 5, they have comparably larger engine_size than group 1 & 2
- Apply ANOVA test for further investigation



Next Steps

- Hypothesis Testing (Further EDA)
- Feature Engineering
  - Categorical Encoding
  - Historical Listing Record Analysis
- Baseline Modeling
  - Logistic Regression - Classification