Machine Learning in Business

MIS710 – A1

Jason Motors Group:

US Market Used Car Pricing Strategy

Name – Himanshi Sachdeva
Student ID - 224850909
Email address – s224850909@deakin.edu.au

Executive Summary

This report investigates the US used car market using real-world data from Craigslist to support Jason Motors Group's (JMG) potential expansion. The dataset includes over 58,000 listings. After rigorous preprocessing, exploratory data analysis, and predictive modelling, a Random Forest Regressor achieved strong results ($R^2 = 0.77$). Key price predictors included age, odometer, condition, and fuel type. Based on insights and performance, this report recommends JMG deploy the model to guide price setting and inventory decisions, improve listing quality, and monitor alternative fuel trends.

1. Business Understanding

JMG is experiencing pricing pressure due to market uncertainty, increasing competition, and shifting demand toward fuel-efficient and electric vehicles. The objective is to apply machine learning to predict car resale prices more accurately to reduce holding time and improve margin.

The key BACCM elements are:

- **Need:** Accurate pricing to reduce holding costs
- **Stakeholders:** JMG sales managers, market researchers
- **Value:** Higher turnover, smarter pricing, better forecasting
- **Solution:** ML-based pricing model trained on Craigslist data
- **Context:** Expanding digital platform in the US market

2. Data Overview & Preprocessing

Initial Dataset Size: 62,946 rows, 18 columns

The dataset contains attributes such as make, model, year, odometer, condition, and more. Key steps in preprocessing included:

- Parsed and standardized Listed_Date
- Created Car_Age from Year
- Filled missing Cylinders with mode
- Replaced missing Region with 'Unknown'
- Removed listings with missing Listed_Price
- Removed outliers in Listed_Price using IQR method
- One-hot encoded categorical variables (resulting in 137 columns total)

Post-cleaning Size: 58,819 rows

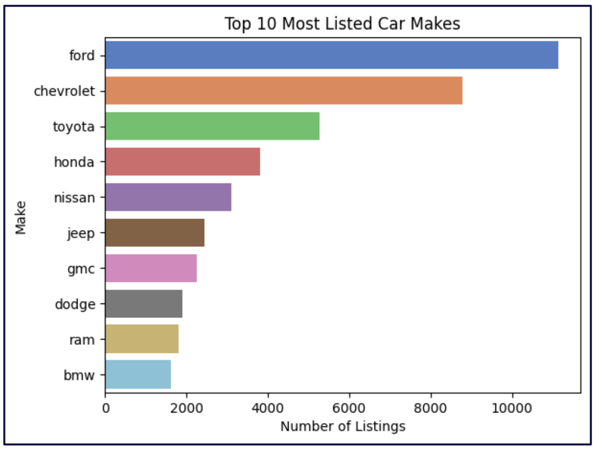
3. Exploratory Data Analysis (EDA)

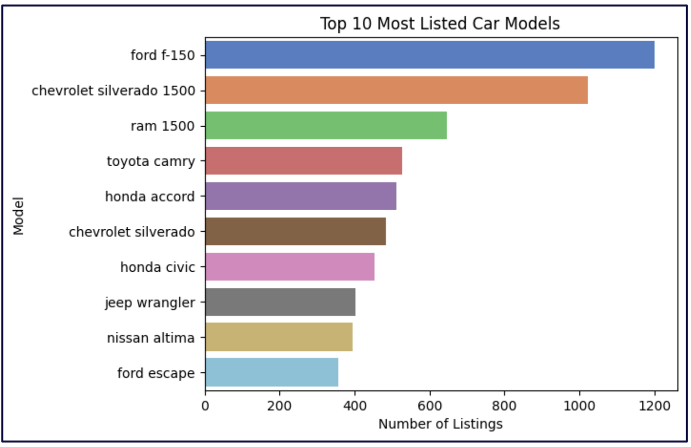
3.1 Overview of Listed Cars

- **Top Makes:** Ford, Chevrolet, Toyota
- **Vehicle Types:** Sedans, SUVs, Trucks
- **Conditions:** Majority are 'excellent' or 'good'
- **Sizes:** Mostly full-size and mid-size

3.2 Popular Makes and Models

- Most listed models include Ford F-150, Toyota Camry, and Chevrolet Silverado

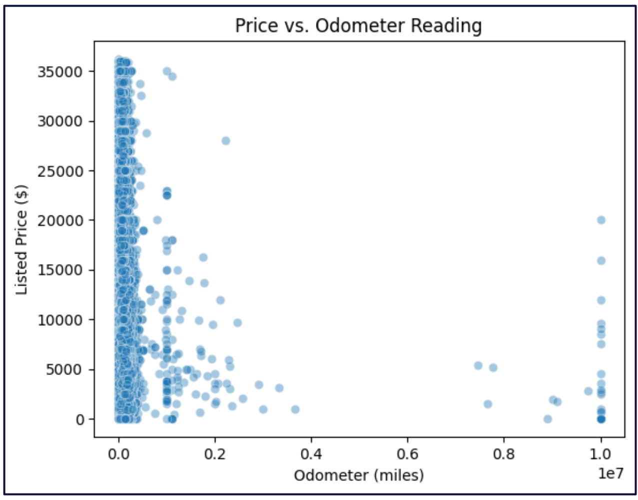


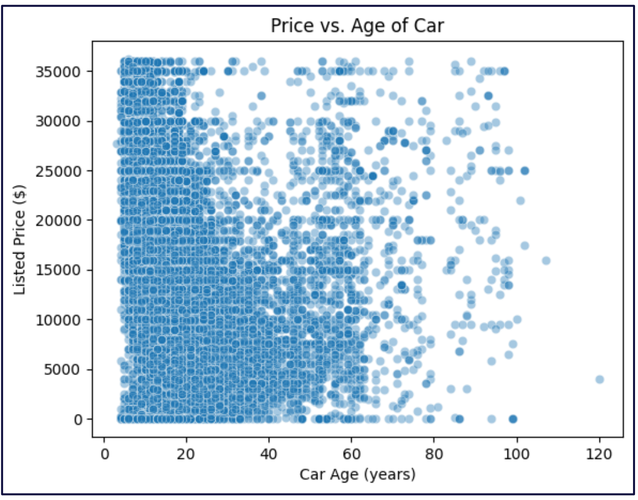


3.3 Price vs Odometer & Age

- Clear negative correlation between price and both age and odometer

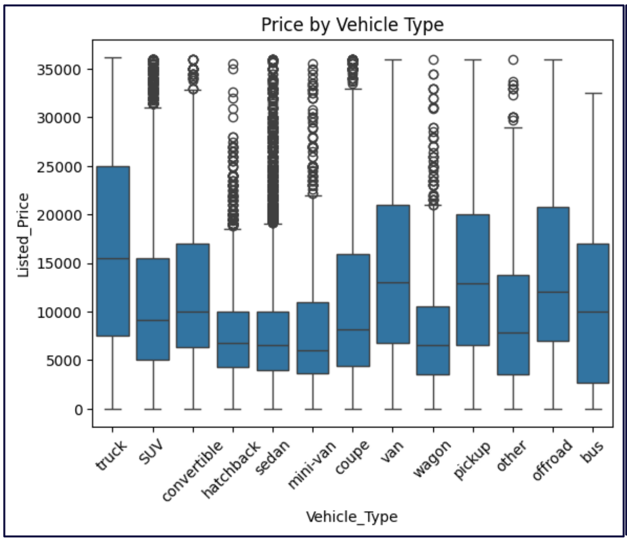
- Outliers present but removed using IQR

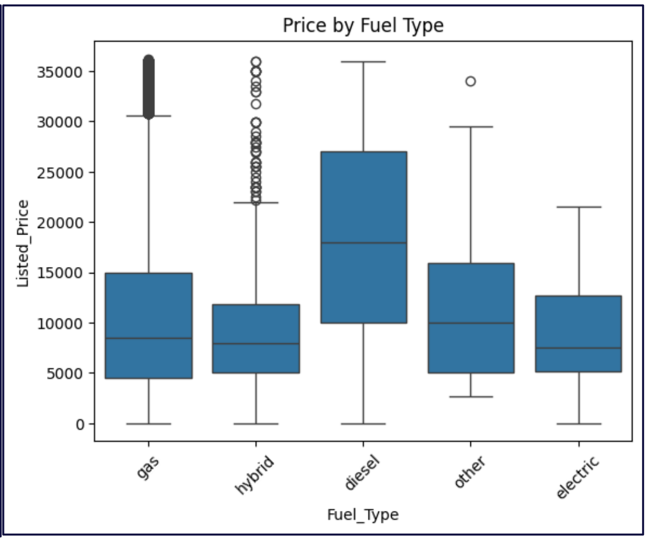


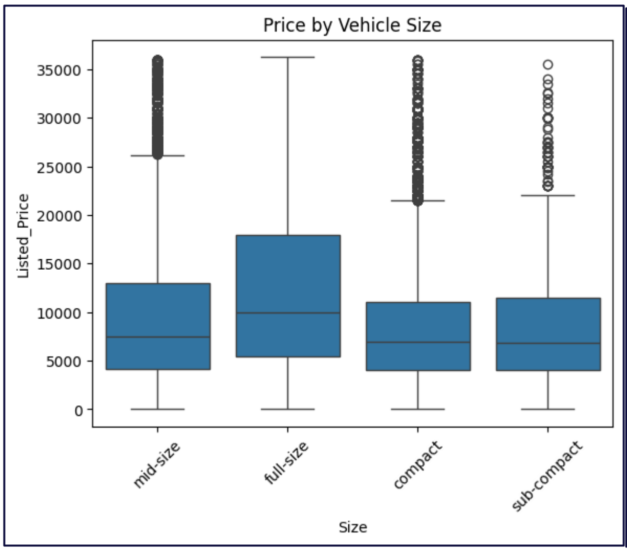


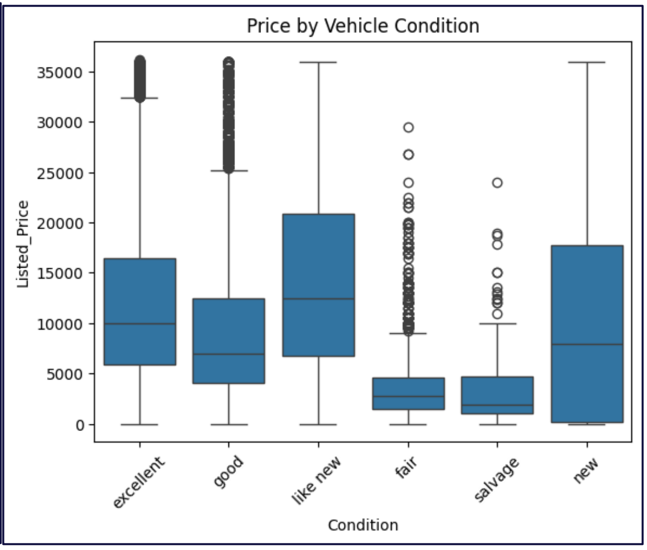
3.4 Price by Features

- Higher prices seen for trucks, diesels, and newer vehicles
- 'Like new' and 'excellent' condition vehicles show highest median prices





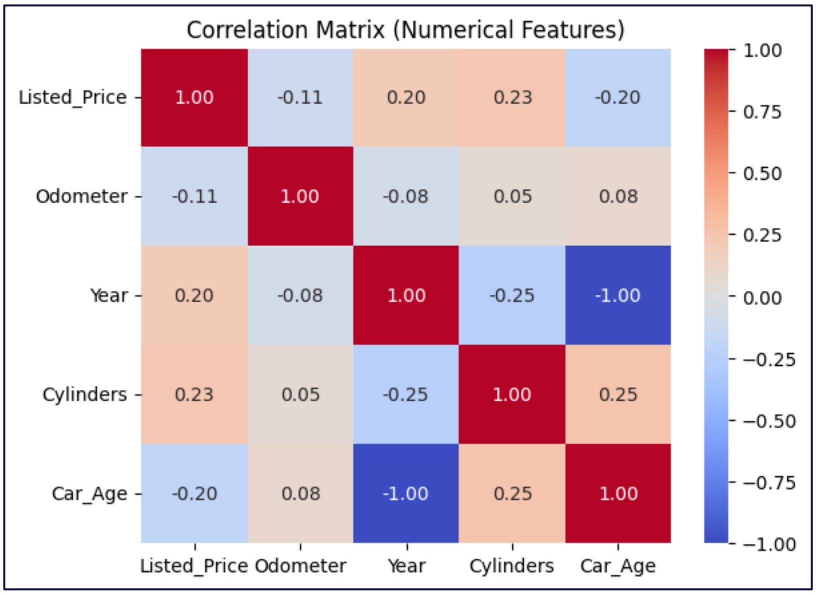


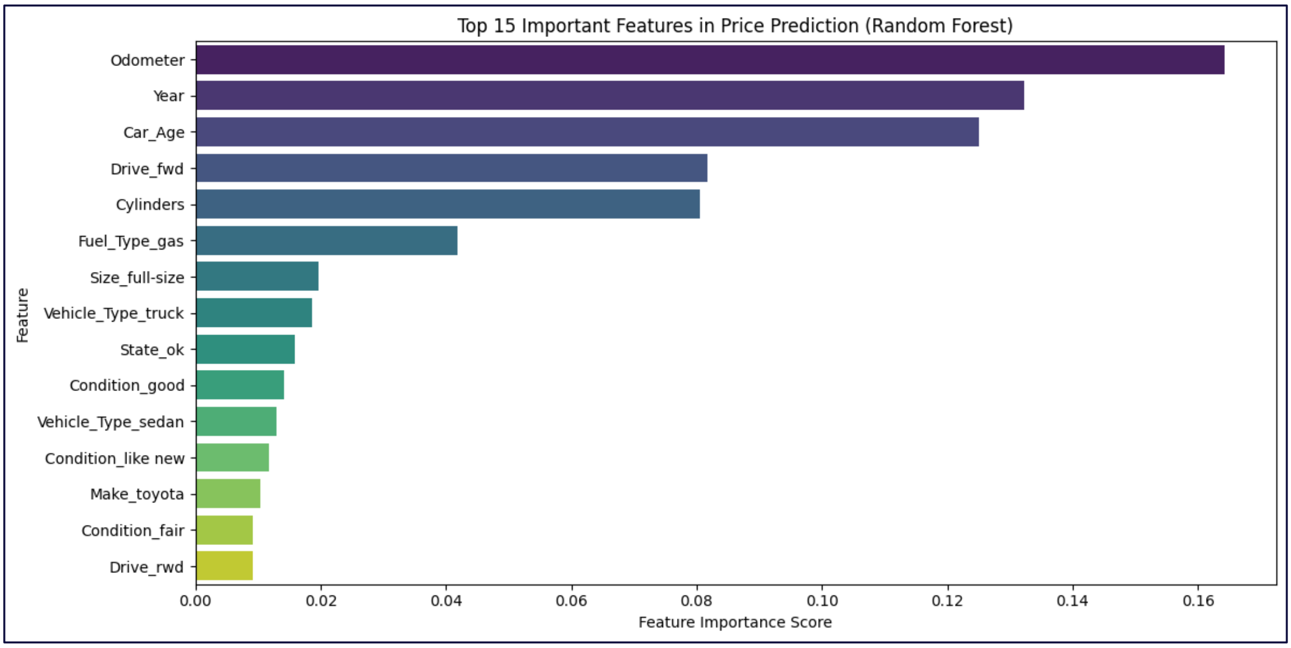


3.5 Additional Price-Affecting Features

- Correlation analysis identified Odometer (-0.43), Car_Age (-0.42), Year (0.43) as key numerical drivers

- Random Forest feature importance showed strong impact of:
    - Odometer, Car_Age, Year, Drive_fwd, Cylinders, Fuel_Type_gas





4. Machine Learning Approach

Two supervised learning models were built:

- **Linear Regression** – interpretable baseline
- **Random Forest Regressor** – nonlinear ensemble method

Train-test split: 80/20 (47,055 training / 11,764 testing) **Feature Count:** 136 features

5. Model Evaluation

MODEL	MAE (↓)	RMSE (↓)	R^2 SCORE (↑)
LINEAR REGRESSION	5164.44	6943.87	0.3269
RANDOM FOREST REGRESSOR	2344.94	4092.28	0.7662

The Random Forest model outperformed the linear baseline and will be used in further implementation. It achieved strong predictive accuracy with lower error metrics across the board.

6. Interpretation: Pros and Cons

Random Forest Pros:

- Captures non-linear patterns
- Handles outliers and multicollinearity
- High accuracy (R^2 = 0.77)

Cons:

- Harder to interpret
- Slower to retrain than linear models

Linear Regression Pros:

- Easy to interpret and deploy
- Fast training

Cons:

- Poor performance due to oversimplified assumptions

7. Business Solution & Recommendations

Recommendation	*Action*	*Benefit*
I. ML-Powered Pricing Tool	Deploy trained Random Forest model	Accurate, automated pricing
2. Focus on Key Drivers	Use mileage, age, type, and condition in price decisions	More competitive and fair valuations
3. Regional Adjustment	Account for price variation across states (e.g., OK, LA)	Enhanced location-specific competitiveness
4. Data Quality Control	Flag unrealistic odometer, year entries	Improves model performance
5. Monitor Alt-Fuel Trends	Track electric and hybrid listings more closely	Prepares JMG for future demand
6. Future Enhancements	Use click data and retrain model regularly	Improves long-term effectiveness and adaptivity

References

1. Drive. (2024, January 19). *Used car prices expected to crash as oversupply hits Australian market*. https://www.drive.com.au/news/used-car-prices-expected-to-crash-as-oversupply-hits-australian-market/
2. IBISWorld. (2024). *Used Car Dealers in the US - Market Research Report*. https://www.ibisworld.com/united-states/industry/used-car-dealers/1004/
3. *Acknowledgement:* Parts of this report were supported using generative AI tools to assist with code generation, spelling correction, and summarisation. The final content has been critically reviewed and edited by the student.