**TORRENS UNIVERSITY AUSTRALIA**

**Introduction to Data Science**

**Assessment 3**

**Case Study: Australia Car Market**

**Luis Alberto Mercado Rodriguez**

**Student ID A00147277**

[LUIS.RODRIGUEZ1@STUDENT.TORRENS.EDU.AU](mailto:LUIS.RODRIGUEZ1@STUDENT.TORRENS.EDU.AU)

**Code:** <https://colab.research.google.com/drive/1OmwxZuYQUSvNuj8W8BzSs019ALAFG3GF?usp=sharing>

**2024**

**Data Science Problem**

The study aims to identify and quantify the factors influencing used car prices in Australia. While traditional factors like mileage and age are well understood, the impact of attributes such as brand, gearbox type, and body design remains unclear. Does the price vary according to the body design, gearbox, brand, or fuel type? This report focuses on providing a data-driven approach to understanding these price determinants.

**Description**  
The dataset used for this study includes 17,952 rows and 16 columns, containing information on used cars in Australia. Key variables include:

* **ID:** Car Code
* **Name:** Car Name
* **Price:** Car Price
* **Brand:** Car Brand
* **Model:** Car Model Year
* **Variant:** Car Variant
* **Series:** Car Series
* **Year:** Year of Manufacture
* **Gearbox:** Gearbox Type
* **Type:** Body Design
* **Fuel:** Fuel Type
* **Status:** Car Condition
* **Kilometers:** Mileage
* **CC:** Engine Capacity
* **Color:** Car Color
* **Seating Capacity:** Seating Capacity

This dataset was sourced from Kaggle, ensuring diversity and reliability. (Australia Car Market Data, 2024)

**Analytics Approach Used**

A structured methodology was involved in dealing with the analysis problem of the Australian used car market. First, we prepared and cleaned the data to guarantee high quality by removing duplicates handling missing values, and encoding categorical variables for consistency. After this, an exploratory data analysis (EDA) was conducted to identify correlations between trends, and key features that influence car prices. We used univariate and bivariate visualizations.

Following this, multiple machine-learning models were implemented to predict car prices based on the features that were identified. We conducted a comparative evaluation of the models used while concentrating on metrics.

**How The Analytics Approach helped to answer the Data Science Problem and the Statistical Method Used.**

The analytics approach solved the problem by identifying the attributes including brand, gearbox type, and body design that influence car prices and allowing accurate predictions with statistical and machine learning techniques. Feature correlations showed which aspects most affected pricing and offered actionable understandings of pricing strategies.

Then, other statistical methods such as correlation analysis and feature scaling were helpful in normalizing and understanding the data and they identified relationships between variables. Machine learning models were evaluated using metrics like R2 and Mean Squared Error (Mean Square Error & R2 Score Clearly Explained, 2018).

**Steps to Analyze the Data**

**1. Data Exploration**

* The functions pd.read\_csv(), df.head(10), and df.info() were utilized to import data and examine the dataset's structure. It facilitates comprehension of variable kinds, names, content overview, and possible problems.

**2. Data Cleaning**

* In the second phase, duplicates were eliminated, missing entries and null values were handled, and irrelevant columns for our case study were dropped. guarantees a clean and pertinent dataset for examination.

**3. Univariate Analysis**

* Applied the use of histograms and bar plots to explore numerical and categorical distributions. Identifies patterns and variable behavior.

**4. Multivariate Analysis**

* To analyze relationships (e.g., Price vs. Brand), boxplots and scatterplots were used. draws attention to dependencies and interactions.

**5. Correlation Analysis**

* To illustrate correlations, use sns.heatmap(). determines the connections between numerical variables. (Correlation in machine learning — All you need to know, 2024)

**6. Hypothesis Testing**

* ANOVA tests on categorical variables were employed in the hypothesis testing process. assesses their statistical importance concerning price. (What is ANOVA and what can I use it for?, 2024)

**7. Feature Engineering**

* To encode and standardize features, LabelEncoder and StandardScaler were utilized. gets data ready for models that use machine learning. (Character encodings, 2023)

**8. Model Training**

* Various Machine learning models like Ada Boost and Random Forest were used with a test of 0.2. (GradientBoosting vs AdaBoost vs XGBoost vs CatBoost vs LightGBM, 2023) uses measures to compare the performance of the model on train and test. (Regression model evaluation metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE, 2023)

**9. Feature Importance**

* UsedRandom Forest to rank features. Identifies the most impactful variables for predictions.

**1. Dataset Overview**

The dataset provides information on cars with the key features mentioned before. Only 65 duplicates were found and eliminated for the study, and there are no missing values.

**Results from Summary Statistics:**

* **Price:**
  + Mean: 36,761 AUD
  + Median: 29,990 AUD lower than the mean, which means that is left skewed
  + Std Dev: 30,334 AUD due to a high standard deviation we can infer that there is a mix of expensive and affordable vehicles
* **Year:**
  + Mean: 2015 indicating that most of the vehicles are recent
  + Median: 2016, is consistent with the mean.
* **Kilometers:**
  + Mean: 103,442 km
  + Std Dev: 80,386 km due to high variability, we can infer that used and nearly-new cars are included.

**2. Relationships Between Variables**

**A screenshot of a chart

Description automatically generatedCorrelation Matrix:**  
The following insights are drawn:

* **Price vs Year:** Positive correlation (0.488). Newer cars tend to be priced higher.
* **Price vs Kilometers:** Negative correlation (-0.447). Higher mileage reduces car value.
* **Price vs CC (Engine Size):** Moderate positive correlation (0.298). Larger engine sizes are associated with higher prices.

**A graph with blue dots and white text

Description automatically generatedA graph of a brand

Description automatically generated3. Price by Brand**

* While luxury brands such as Porsche, Lamborghini, Mercedes, Audi and BMW have significantly higher prices the last three are the most popular brands.
* Toyota and Holden dominate in the count but have lower average prices.

**A graph showing body type popularity

Description automatically generated4. Price by Type**

A graph of different colored objects

Description automatically generated with medium confidence

* Wagons, Sedans, and Hatchbacks are most common, but Wagons have the highest price range.
* Specialized types often command higher prices.

**5. Price by Fuel Type**

A graph of a graph with numbers and lines

Description automatically generated with medium confidenceA graph of fuel type distribution

Description automatically generated

* Diesel and premium unleaded cars tend to have higher prices.
* Hybrid (petrol/electric) vehicles fall into the middle range but are less common.

**A graph of different colors

Description automatically generated6. Gearbox Insights**

A graph with blue dots

Description automatically generated

* Automatic cars dominate the market (85%).
* Very few cars have AWD or specialized gearbox types.
* Some manual cars have uncommon prices, this relationship may be related to the fact that some luxury brands offer manual options for their cars.
* Vehicles different from Automatic and Manual have lower prices.

A black screen with white text

Description automatically generated**7. ANOVA Tests**

* **Brand:** Luxury brands significantly affect price.
* **Type:** Wagon or Hatchback types affect price.
* **Status:** New cars in stock fetch much higher prices than used/demo models.
* **Gearbox:** Indicate significant price differences between transmission types

A screenshot of a graph

Description automatically generated**8. Model Performance Comparison**

* **Best Models:**
  + ExtraTreesRegressor and CatBoostRegressor: Highest R² score on the test set (0.847) and lowest mean squared error.

For enhancing the models further enhancements should be made.

**A graph with blue bars

Description automatically generated9. Importance Variation**

* Brand accounts for approximately ~8% of price variation
* Fuel type accounts for approximately ~7% of price variation
* Body type accounts for ~3% of price variation
* Gearbox accounts for ~0.6% of price variation

**Conclusion**

After more conventional criteria like age and mileage, brand reputation (varies by 8%) and fuel type (varies by 7%) are the top pricing predictors, according to an analysis of 17,952 used automobiles in Australia. Even though luxury manufacturers are more expensive, Toyota and Holden control a large portion of the market, with 85% of listings having automatic gearboxes. Machine learning algorithms give market participants important insights by showing that transmission type (0.6%) and body design (3%) have less of an effect on pricing than previously thought. ***(1083 WORDS)***

**Glossary**

**Correlation:** Correlation is a key statistical concept that researchers employ to analyze connections within their data. It helps us to Understand the Relationship Between Variables(Correlation in machine learning — All you need to know, 2024)

**Anova:** Analysis of variance (ANOVA) is a statistical test used to assess the difference between the means of more than two groups. At its core, ANOVA allows you to simultaneously compare arithmetic means across groups. You can determine whether the differences observed are due to random chance or if they reflect genuine, meaningful differences. (What is ANOVA and what can I use it for? , 2024)

**R-Squared:** It measures the proportion of the total variation in the dependent variable that is captured by the model. (Regression model evaluation metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE, 2023)

**Mean Squared Error:** This is the average of the squared differences between the predicted and actual values. It gives more weight to larger differences and is particularly useful when we have unexpected values that we want to take into account. (Regression model evaluation metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE, 2023)

**Encoding:** Character encodings are specific sets of rules for mapping from raw binary byte strings (that look like this: 0110100001101001) to characters that make up human-readable text (like "hi"). (Character encodings, 2023)

**References**

* Australia car market data. (2024, September 23). Kaggle. <https://www.kaggle.com/datasets/lainguyn123/australia-car-market-data/data>
* Rowe, W. (n.d.). Mean square error & R2 score clearly explained. BMC Blogs. <https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/>
* Ashraf, A. (2024b, February 20). Correlation in machine learning — All you need to know. Medium. <https://medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9>
* What is ANOVA and what can I use it for? | Qualtrics AU. (2024, March 7). Qualtrics. <https://www.qualtrics.com/en-au/experience-management/research/anova/>
* GeeksforGeeks. (2023, May 29). GradientBoosting vs AdaBoost vs XGBoost vs CatBoost vs LightGBM. GeeksforGeeks. <https://www.geeksforgeeks.org/gradientboosting-vs-adaboost-vs-xgboost-vs-catboost-vs-lightgbm/>
* Wohlwend, B. (2023, August 12). Regression model evaluation metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE. Medium. <https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3#:~:text=Mean%20Squared%20Error%20(MSE)%20is,the%20actual%20and%20predicted%20values>.
* Alexisbcook. (2023, April 20). Character encodings. Kaggle. <https://www.kaggle.com/code/alexisbcook/character-encodings>