Automobile Imports-85 Dataset

Statistical Methods Project Overview

March 16, 2025

Introduction

- In this project, we are exploring the **Automobile Imports-85 Dataset**.
- The dataset contains various attributes of automobiles including technical specifications, performance metrics, and price.
- Our goals:
  - Provide a comprehensive overview of the dataset.
  - Understand the theoretical aspects of the methods we plan to use.
  - Propose analyses based on hypothesis testing and predictive modeling.

General Description of the Dataset

- The dataset is sourced from the UCI Machine Learning Repository.
- It contains **205** instances representing different automobiles. Some records are incomplete.
- The attributes include both categorical and numerical variables. Examples include:
  - **Symboling**: A risk rating for the automobile (ranging from -3 for very safe to +3 for very risky).
  - **Normalized Losses**: A normalized measure of the insurance losses, which indicates repair costs.
  - **Make**: The manufacturer of the automobile (e.g., alfa-romero, audi, bmw).
  - **Fuel Type** and **Aspiration**: Attributes describing the fuel used and engine aspiration (e.g., standard, turbo).
  - **Engine Size** and **Horsepower**: Key performance metrics.
  - **Price**: Often used as a target variable for regression modeling.

Detailed Summary of Attributes

- **Symboling**:
  - Risk factor with values typically ranging from -3 (low risk) to +3 (high risk).
- **Normalized Losses**:
  - Represents relative insurance losses; higher values imply higher repair/maintenance costs.
- **Make**:
  - The car manufacturer, which may influence design, performance, and pricing.
- **Fuel Type** and **Aspiration**:
  - Fuel type (e.g., gas, diesel) and whether the engine is naturally aspirated or turbocharged.
- **Engine Attributes**:
  - Includes engine size (in cubic centimeters), horsepower, and other performance metrics.
- **Performance Metrics**:
  - Such as city and highway MPG, which reflect fuel efficiency.
- **Price**:
  - A key continuous variable often used for regression analysis.

Expected Relationships Between Attributes

Based on automotive engineering and market dynamics, we expect:

- Higher **engine size** and **horsepower** to correlate with higher **prices**.
- **Fuel type** and **aspiration** may influence fuel efficiency (city/highway MPG) and overall performance.
- **Symboling** (risk rating) might be associated with insurance costs and repair expenses.
- Associations among categorical variables (e.g., fuel type, body style) can be explored using chi-square tests.

Hypothesis Testing: Detailed Overview

Our hypothesis testing will focus on evaluating associations between variables:

- **Null Hypothesis (H_0)**: Assumes no association between the variables under investigation (e.g., "Engine size is independent of price").
- **Alternative Hypothesis (H_1)**: Assumes there is a statistically significant association (e.g., "Engine size is associated with price").
- **Testing Procedure:**
  - For categorical variables, we will use the **Chi-square test of independence**.
  - For numerical relationships, correlation tests or regression analyses will be used.
- **Statistical Significance:**
  - A p-value less than $\alpha = 0.05$ will lead us to reject the null hypothesis.
  - Effect sizes, such as **Cramér's V**, will be computed to quantify the strength of associations.

Predictive Modeling: Decision Tree Analysis

Decision Trees will be used for classification and regression:

- **Algorithm:**
  - We plan to use the CART (Classification and Regression Trees) algorithm.
  - The tree splits data based on impurity measures (e.g., Gini index or information gain) for classification, and variance reduction for regression.
- **Model Building:**
  - Data will be split into training and testing subsets.
  - Cross-validation techniques will be used to prevent overfitting.
- **Interpretability:**
  - Decision trees provide intuitive rules that help identify the most influential attributes.

Predictive Modeling: Regression Analysis

Regression Analysis will help us predict continuous outcomes such as **price**:

- **Linear Regression:**
  - Models the relationship between one or more predictor variables and a continuous target variable.
  - Assumes linearity between predictors and the target.
- **Model Evaluation:**
  - Performance metrics include R-squared, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error).
  - Diagnostic plots will be used to assess assumptions (e.g., normality of residuals, homoscedasticity).
- **Extensions:**
  - If the linear model is insufficient, we may explore polynomial regression or regularization techniques (e.g., Ridge, Lasso).

Proposal of Analyses

Our comprehensive analysis plan includes:

1. **Data Cleaning and Preparation:**
   - Handle missing values and encode categorical variables appropriately.
2. **Exploratory Data Analysis (EDA):**
   - Compute frequency distributions and visualize each attribute using bar plots, histograms, and scatter plots.
3. **Inferential Statistics and Hypothesis Testing:**
   - Use chi-square tests for categorical variables and correlation tests for numerical variables.
   - Formulate null and alternative hypotheses (e.g., "Engine size is associated with price").
4. **Predictive Modeling:**
   - **Decision Trees:** Build classification/regression trees using CART with cross-validation.
   - **Regression Analysis:** Develop linear or polynomial regression models to predict price, evaluating model assumptions and performance.

Conclusion and Future Work

- Our study aims to identify the key factors that influence automobile characteristics and pricing.
- Future work may include:
  - Incorporating ensemble methods or other advanced modeling techniques.
  - A deeper analysis of variable interactions and model refinements.
- Our findings will be compared with established automotive market theories to validate the insights.

References

Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository: Automobile Data Set*. Available at: `https://archive.ics.uci.edu/ml/datasets/Automobile` (Accessed: March 13, 2025).

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.