**Analysis of Car Sale Advertisements**

***Unveiling Trends and Patterns in the Automotive Market***

**University of North Texas**

**CSCE 5310: Methods In Empirical Analysis**

Dr. Mark Albert

|  |
| --- |
| 1.Lakshmi Prasanna Valdas |
| 2.Dedeepya Vaddi |
| 3.Sekhar Reddy Kandula |
| 4.Rajeswari Amrutha B |

**TEAM MEMBERS:**

**Abstract:**

The process of buying an automobile now involves a lot of internet research and discovery, condition assessment, finance, and transaction. Additionally, several variables are in play to encourage higher consumer conversion. The price of the automobile, its mileage, and its condition are some of the most important aspects that needs to be taken into consideration.

Keeping all the automobiles' listing prices competitive without cutting into profits is one of the toughest challenges facing new competitors. The company must understand what drives market pricing for automobiles. Therefore, a thorough investigation will aid management in comprehending how pricing fluctuate in the market for used cars.

To identify trends and patterns that impact the automobile market, this research uses statistical approaches to dive into the dynamic world of vehicle sale marketing data. The project makes use of a large dataset that contains a variety of information about automobile commercials, such as characteristics, price, and demographic information. This investigation aims to provide answers to important issues about customer behavior, market dynamics, and the efficacy of marketing tactics in the automobile sector by using both advanced statistical analyses and preliminary exploratory data analysis.

**Data Description and Statistical tests:**

The dataset that was obtained from Otomoto.pl is a large collection of vehicle sale ads that includes 208,304 entries and 25 different types of data attributes. Along with maybe more subtle information like brand, model, and fuel type, these features probably contain standard aspects like price, mileage, power, and emissions. Analysts have a strong basis for investigating the automobile industry and identifying customer preferences with such a wealth of diverse and comprehensive information at their disposal. Every record contains essential information on specific vehicle listings, giving users a comprehensive picture of the cars that are listed on the site. Through a thorough analysis of these data points, scholars may get crucial information into pricing patterns, geographical discrepancies, and the changing dynamics of the automobile sector.

This dataset might be used to draw the following conclusions. They are:

* Price Analysis: We may look at the variables like age, brand, condition, and mileage that affect how much used automobiles cost. One way to model the link between the price and different independent variables is to utilize regression analysis.
* Brand Preference Analysis: We may learn about customer preferences and the relative popularity of various vehicle brands by examining the dataset's brand distribution.
* Analysis of Mileage and Age: We may investigate the link between an automobile's age and mileage and how it affects the value and cost of used automobiles.
* Emissions and Environmental Impact: The collection includes statistics on CO2 emissions, which may be examined to determine how various automobile models affect the environment and if there is a market for more environmentally friendly vehicles.
* Geographic Analysis: We might look at regional variations in vehicle prices, preferences, and market dynamics provided the dataset has location data.

A mix of parametric and non-parametric tests may be taken into consideration, depending on the features of the dataset and the analytic goals.

**Parametric Tests:**

1. one-sample t-test: To compare the mean mileage of the automobiles in the dataset with a hypothetical figure, such 50,000 kilometers.
2. Independent t-test: It is also possible to compare the mean cost of automobiles with diesel engines with cars with gasoline.
3. Analysis of Variance (ANOVA): Further evaluation of the variations in mileage across automobiles from various areas may be ANOVA.
4. Pearson correlation: It may be used to investigate the linear relationship between variables such as miles and price.

**Non-Parametric tests:**

1. Mann-Whitney U test: It is appropriate for non-parametric testing for comparing the medians of two independent groups, such as the cost of diesel-powered automobiles vs gasoline-powered cars.
2. Wilcoxon signed-rank test: The Wilcoxon signed-rank test can be utilized to compare the medians of the same variable under two different conditions, such as comparing the city\_mpg of cars with automatic transmissions to those with manual transmissions.
3. Kruskal-Walli’s test: Comparing the medians of more than two independent groups is also possible using the Kruskal-Wallis test, which may be used to examine mileage variations across vehicles from different geographic areas.
4. Spearman correlation: The strength and direction of a monotonic relationship between variables, like the correlation between automobile age and price, may also be evaluated using it.

**Resources and Related Projects:**

1. **Otomoto Website** (<https://www.otomoto.pl/>): Otomoto is the largest online platform for buying and selling cars in Poland. It hosts a massive database of vehicles from various dealers and private sellers, connecting millions of potential buyers with a wide range of options.
2. **GitHub Repository for Web Scraping** (<https://github.com/bpieniak/otomoto-webscrape>): This GitHub repository provides Python scripts for web scraping data from the Otomoto website. It allows users to extract valuable information about car listings, including details like price, mileage, brand, model, and more.
3. **Kaggle Dataset** (<https://www.kaggle.com/datasets/bartoszpieniak/poland-cars-for-sale-dataset/data>): This Kaggle dataset contains the scraped car sale advertisement data from the Otomoto website, which we will be using for our project. It provides a comprehensive collection of features and observations, enabling us to perform various analyses and build predictive models.

In our project, we aim to go beyond simply utilizing the provided resources. We will conduct extensive exploratory data analysis, feature engineering, and model selection processes to gain insights into the factors influencing car prices and develop accurate predictive models. Additionally, we will explore other aspects of the data, such as brand preferences, regional differences, and environmental impact, to provide a holistic understanding of the automotive market in Poland.

Related Tutorials and Approaches:

1. **"Predicting Used Car Prices with Linear Regression"** by Dataquest (freely available): <https://community.dataquest.io/t/predicting-car-prices-m155/554907>
   * This tutorial covers using linear regression to predict used car prices based on various features, similar to our project.
   * It provides a step-by-step guide on data preprocessing, feature selection, model training, and evaluation.
   * While this tutorial focuses primarily on linear regression, we plan to explore additional machine learning techniques, such as decision trees, random forests, and gradient boosting, to improve model performance.
2. **"Used Car Price Prediction"** by Kaggle User (freely available, open-source): <https://www.kaggle.com/code/manishkr1754/used-car-price-prediction>
   * This project explores different machine learning models, including Random Forest and XGBoost, to predict used car prices.
   * It showcases the application of ensemble techniques and feature importance analysis.
   * Our approach will differ in terms of the dataset used (focused on the Polish market) and the inclusion of additional analyses, such as brand preferences and environmental impact.

By combining the provided resources with insights from related tutorials and approaches, we aim to deliver a comprehensive analysis of the car sales data, tailored to the Polish market, while offering unique perspectives and insights beyond existing work.

**Data Specification:**

The dataset we are working with is a features x observations matrix, where each row represents a car sale advertisement, and the columns represent various features of the advertised cars. The dataset contains 208,304 observations (after removing duplicates) and 25 features.

The features include:

1. **Numerical Features:**

* **price: The listed price of the car (continuous).**
* **year\_of\_production: The year the car was manufactured (discrete).**
* **mileage: The mileage of the car in kilometers (continuous).**
* **engine\_capacity: The engine capacity or displacement of the car (continuous).**
* **power: The engine power of the car, likely measured in horsepower or kilowatts (continuous).**
* **co2\_emission: The carbon dioxide emission level of the car (continuous).**

1. **Categorical Features:**

* **brand: The brand or make of the car (e.g., Toyota, Honda, Ford).**
* **model: The specific model of the car (e.g., Corolla, Civic, Mustang).**
* **fuel\_type: The type of fuel used by the car (e.g., gasoline, diesel, electric).**
* **transmission: The transmission type of the car (e.g., manual, automatic).**
* **body\_type: The body style of the car (e.g., sedan, hatchback, SUV).**
* **condition: The condition of the car (e.g., new, used).**
* **location: The geographical location or region where the car is advertised.**

**Design and Milestones:**

The project will be implemented using Python and the following libraries and frameworks:

1. **Pandas**: A powerful data manipulation and analysis library for Python. We will use Pandas for loading the dataset, handling missing values, and performing data preprocessing tasks.
2. **NumPy**: A fundamental library for scientific computing in Python. NumPy will be used for numerical operations and array manipulation throughout the project.
3. **Matplotlib** and **Seaborn**: Popular data visualization libraries in Python. We will utilize these libraries for creating various plots and visualizations during the exploratory data analysis (EDA) phase.
4. **Scikit-learn**: A machine learning library for Python. Scikit-learn will be used for feature selection techniques, building and training different regression models, and evaluating their performance.
5. **Statsmodels**: A Python library for statistical modeling and data analysis. We will employ Statsmodels for conducting various statistical tests, such as t-tests, ANOVA, and non-parametric tests.
6. **Jupyter Notebook**: An open-source web application that allows us to create and share documents containing live code, visualizations, and narrative text. We will use Jupyter Notebook as our primary development environment for this project.

**Models and Techniques:**

1. **Linear Regression**: As a baseline model, we will train a linear regression model to predict car prices based on the selected features. Linear regression assumes a linear relationship between the independent variables (features) and the dependent variable (car price).
2. **Feature Selection Techniques**:
   * **Correlation Analysis**: We will calculate correlation coefficients (e.g., Pearson's or Spearman's) between the features and the target variable (car price) to identify highly correlated features.
   * **Mutual Information**: We will compute the mutual information between each feature and the target variable to assess the relevance and importance of features.
   * **Recursive Feature Elimination (RFE)**: RFE is a technique that recursively removes features with the least importance, based on a specified machine learning model (e.g., linear regression or random forest).
3. **Model Evaluation and Selection**:
   * **Cross-Validation**: We will perform k-fold cross-validation to evaluate the performance of our models and minimize the risk of overfitting. Relevant metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared will be computed.
   * **Holdout Validation**: Additionally, we will split the dataset into training and test sets and evaluate the performance of our models on the unseen test set.
4. **Hyperparameter Tuning**:
   * **Grid Search**: We will employ grid search to systematically search for the optimal hyperparameters of our models by evaluating their performance across a specified range of values.
   * **Random Search**: Alternatively, we may use random search, which randomly samples hyperparameter values from a specified distribution, to find the best combination of hyperparameters.

**Milestones:**

**Here are some verifiable milestones for this project:**

**1. Data Pre-processing:**

* **Milestone 1: Loaded and clean the dataset.**
* **Milestone 2: Handled missing data and outliers.**
* **Milestone 3: Performed feature engineering.**

**2. Exploratory Data Analysis (EDA):**

* **Milestone 1: Performed statistical distribution analysis.**
* **Milestone 2: Performed categorical data analysis.**
* **Milestone 3: Performed numerical data analysis.**
* **Milestone 4: Perform data distribution analysis of various features.**

**3. Statistical tests:**

* **Milestone 1: Performed t-tests for means comparison.**
* **Milestone 2: Performed ANOVA for means comparison.**
* **Milestone 3: Performed Mann-Whitney U test for distribution comparison.**
* **Milestone 4: Performed Kruskal-Wallis test for distribution comparison.**

**4. Feature Selection:**

* **Milestone 1: Performed feature selection using correlation analysis.**
* **Milestone 2: Performed feature selection using mutual information.**
* **Milestone 3: Performed feature selection using recursive feature elimination.**

**5. Model Selection:**

* **Milestone 1: Performed model training and hyperparameter tuning.**
* **Milestone 2: Performed model evaluation using cross-validation.**
* **Milestone 3: Performed model evaluation using holdout validation.**

**Dataset**- The dataset for this project is taken from the Kaggle.

https://www.kaggle.com/datasets/bartoszpieniak/poland-cars-for-sale-dataset/data -

This Kaggle page provides the scrapped car sale advertisement data for public, which is used in the current project.

*Input-Output pair: The dataset is obtained using web scrapping car advertisement data from Otomoto.pl [1] site using the code in [2].*

The dataset contains 25 data features and 208,304 records. [3]

A screenshot of a computer

Description automatically generated

**1. Data Collection and Preprocessing**

https://www.kaggle.com/datasets/bartoszpieniak/poland-cars-for-sale-dataset/data - This Kaggle page provides the scrapped car sale advertisement data for public, which is used in the current project.

***Importing the data-***

***A screenshot of a computer

Description automatically generated***

***Understanding the data-***

***A screenshot of a computer

Description automatically generated***

***Removing duplicates from the dataset-***

***A screenshot of a computer

Description automatically generated***

**2. Exploratory Data Analysis (EDA)**

EDA will be used to analyze the data, display it, create density charts, distribute charges, find outliers, and investigate correlations between the data's components.

***Statistical Distribution-*** The distinct value in a statistical distribution indicates how many different values there are in an attribute. Skewness provides the attribute's mode via contrasting with by determining if the distribution is skewed to the left or right, we may determine whether it is flat or peaks in relation to a normal distribution.

A screenshot of a computer

Description automatically generated

***Histogram for numerical columns-***

A graph of a person with histogram

Description automatically generated

A graph with a bar

Description automatically generated

A graph of a person with histogram

Description automatically generated

A graph of mileage

Description automatically generated

A graph of power

Description automatically generated

A graph of a column

Description automatically generated

A graph of co2 emissions

Description automatically generated

A graph with numbers and text

Description automatically generated

***Categorical Data Analysis-***

***A close-up of a text

Description automatically generated***

***Null value analysis-***

***A white background with black text

Description automatically generated***

***A purple and yellow lines with white text

Description automatically generated***

***Numerical Data Analysis-***

***A screenshot of a computer

Description automatically generated***

***A screenshot of a computer

Description automatically generated***

**3. Statistical tests**

Two major types of statistical tests—parametric and non-parametric—are utilized to examine data in various contexts according to presumptions on the distribution of the data and the degree of variable measurement.

1. ***Parametric tests-***

* **One-sample t-test:** Compares the mean of a single sample to a known value.
* **Independent t-test:** Compares the means of two independent groups.
* **ANOVA (Analysis of Variance):** Compares the means of more than two independent groups.
* **Pearson correlation:** Measures the linear relationship between two continuous variables.

1. **Non-Parametric tests-**

* **Wilcoxon signed-rank test**: Compares the median of a single sample to a known value.
* **Mann-Whitney U test**: Compares the medians of two independent groups.
* **Kruskal-Wallis test**: Compares the medians of more than two independent groups.
* **Spearman correlation**: Measures the monotonic relationship between two continuous variables.

**4. Feature Selection**

***Finding correlation between variables-***

**A white background with colorful lines

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

***Performing feature selection using mutual information-***

***A white screen with black text

Description automatically generated***

***Performing feature selection using recursive feature elimination-***

***A screenshot of a computer program

Description automatically generated***

**5. Model Selection**

In Model Selection, we employed three model evaluation milestones:

***Training model and hyper-parameter tuning-***

To train the model and tune hyperparameters, we employed GridSearchCV from the sklearn library. We headed by preprocessing the data to determine the feature matrix X and the target variable y. StandardScaler is applied for scaling the features, and we divided the data into training and testing sets by train\_test\_split.

A pipeline is created that involves the Linear Regression model and feature scaling.

***A white screen with text

Description automatically generated***

We performed Hyperparameter tuning with GridSearchCV by the indicated parameters for 'fit\_intercept' and 'positive'. To find the best composition of hyperparameters a grid search is carried out on the training data.

***A screenshot of a computer code

Description automatically generated***

***Model evaluation using cross-validation-***

We implemented Cross-validation with 'cross\_val\_score` to assess the model performance and prevent overfitting. To determine the effectiveness of the model we utilized k-fold cross-validation (cv=5). We divided the dataset into k=5 folds and trained the model using four folds and validation was performed on the leftover fold. The whole process occurs k times, with every fold taking shifts as the test set once. To estimate the performance of the model we employed the negative mean squared error as the evaluation measure. We obtained an accurate estimation of the models’s performance by computing the mean of the RMSE scores from all iterations.

***A screenshot of a computer

Description automatically generated***

******

***Model evaluation using fold-out validation-***

We divided the dataset into two sets: training and testing. About 70-80% of the data is used for training, and the rest is kept for testing. We utilized a training set to train the model and then examined its efficiency on the test set by executing specific performance measures. Upon evaluating each of these stages, we obtained an extensive comprehension of the model's predictions for car prices.

***A screenshot of a computer program

Description automatically generated***

***A close-up of a person's face

Description automatically generated***

**Repository / Archive:**

* The project repository is hosted on GitHub and can be accessed at the following link: <https://github.com/Sekhar0799/Analysis-of-car-sale-advertisements>

The repository includes:

* A README file with project overview, installation instructions, and usage guidelines.