D1 (Uku Renek Kronbergs, Janser-Jax Lang)

Used car price prediction project 08/12/2024

Repository: https://github.com/UkuRenekKronbergs/IDS_project

This project aims to develop a model for predicting used car prices based on key attributes. We will analyze a dataset of 4009 used cars, identifying the most influential factors affecting price and visualizing depreciation trends. The successful model will have a mean absolute error (MAE) of less than 15% compared to actual market prices. Insights will be presented in a clear and actionable format, aiding car buyers, sellers, and dealers in making informed decisions.

Business Understanding

- **Background:** The used car market is dynamic, influenced by various factors. Accurate price estimation benefits both buyers and sellers. Understanding depreciation trends helps consumers make informed choices and supports businesses in setting pricing strategies.
- **Business Goals:**
  - Develop a predictive model for used car prices.
  - Identify the most influential factors affecting car prices.
  - Analyze average depreciation trends for different car brands and models.
- **Business Success Criteria:**
  - Model accuracy with MAE less than 15% compared to market prices.
  - Clear visualization of depreciation trends.

Inventory of Resources

- **Dataset:** 4009 records with 9 attributes (brand, model, year, mileage, fuel type, engine type, transmission, accident history, clean title).
- **Software and Tools:** Python (libraries like pandas, scikit-learn, matplotlib, seaborn), Jupyter Notebook, cloud-based collaboration tools.
- **Team Skills:** Familiarity with machine learning, data visualization, and statistical analysis.

- **Requirements:**
    - Clean and preprocess the dataset.
- **Assumptions:**
    - Dataset is representative of the broader used car market.
    - Features capture price-influencing variables.
- **Constraints:**
    - Limited dataset size (4009 records).

Risks and Contingencies

- **Risk:** Noisy or irrelevant data.
    - **Contingency:** Robust preprocessing and feature selection.
- **Risk:** The model may not achieve the desired level of accuracy.
    - **Contingency:** Explore alternative algorithms and adjust model settings to improve performance.

Terminology

- **Depreciation trend:** How car loses value over time.
- **Clean Title Status:** Indicates no major incidents impacting ownership.

Costs and Benefits

- **Costs:** Time investment.
- **Benefits:** Accurate price prediction saves money, improves transaction efficiency, provides valuable market insights.

Data-Mining Goals

- Train a regression model to predict car prices.
- Conduct feature importance analysis.
- Compute and visualize depreciation trends.

Data-Mining Success Criteria

- Model performance metrics meet thresholds.
- Feature importance aligns with domain knowledge.
- Depreciation trends are clear and offer practical insights.

Data Understanding

Gathering Data

- Required features: Brand & Model, Model Year, Mileage, Fuel Type, Engine Type, Transmission, Accident History, Clean Title, Price.
- Data Availability: Used Car Price Prediction Dataset (Cars.com) with 4009 records and all required features. Potential missing or noisy data in some fields.
- Selection Criteria:
  - Include cars with valid data for model year, mileage, and price.
  - Focus on mainstream fuel types.
  - Filter out records with implausible values.

Describing Data

- Size: 4009 records, 10 attributes.
- Data Types:
  - Categorical: Brand & Model, Fuel Type, Engine Type, Transmission, Accident History, Clean Title (optional: Exterior & Interior Colors).
  - Numerical: Model Year, Mileage, Price.

Exploring Data

- **Depreciation Analysis:** Cars lose value as they age, with significant drops in price after 5-10 years.
- **Mileage Impact:** Higher mileage correlates with lower prices, with a sharp decline after 150,000 km.
- **Fuel Type Influence:** Electric and hybrid cars generally command higher prices than gasoline and diesel vehicles of similar age.
- **Transmission Type:** Automatic cars show higher average prices, reflecting consumer preference in certain markets.
- **Accident History:** Cars with prior accidents are typically priced 15-20% lower than clean title vehicles.

Initial Visualizations

- Scatter Plot: Price vs. Model Year (colored by fuel type)
- Box Plot: Price distribution by transmission type
- Histogram: Mileage distribution across the dataset

Completeness

- Missing values are noted in accident history (12%) and clean title (8%)
- Some records have zero or null values for milage and price

Consistency

- Price: Outliers include extremely high prices for older vehicles, which need further review.

Accuracy

- Cross-referenced price data shows most values are plausible, but some outliers suggest atypical market listings.

Project Timeline

Task	**Description**	**Hours**
1. Data Collection and Preprocessing	Cleaning and preparing the dataset. Handle missing values, outliers, and normalize data.	6
2. Exploratory Data Analysis (EDA)	Perform detailed analysis of trends and relationships in the data, using visualizations. Identify key features.	6
3. Model Development	Build predictive models (e.g., linear regression, random forest). Train and validate models.	6
4. Feature Analysis	Analyze feature importance to determine which factors influence car prices the most.	4
5. Depreciation trend Calculation	Calculate depreciation trends for different brands and models based on age and price trends.	2
6. Reporting and Documentation	Compile findings into a report and presentation. Include visuals and insights.	6

Team Contribution

Each team member will contribute equally across all tasks, with an estimated workload of 30 hours per person.

Methods and Tools

- Programming Languages: Python (primary), libraries like Pandas, NumPy, Scikit-learn, and Matplotlib/Seaborn for EDA.
- Data Preprocessing Tools: Pandas, Scikit-learn for normalization and encoding.
- Visualization: Matplotlib, Seaborn for insights; Plotly for interactive visualizations.
- Modeling: Linear regression, Random Forest, Gradient Boosting models (using Scikit-learn).
- Report Tools: MS Word, PowerPoint for documentation and presentation.

Estimated project duration: 14 days (assuming we can focus and cooperate).

Next Steps

1. **Data Cleaning and Preprocessing:** Handle outliers, and normalize data.
2. **Exploratory Data Analysis (EDA):** Visualize data distributions, correlations, and trends.
3. **Feature Engineering:** Create new features if necessary (e.g., car age, fuel efficiency).
4. **Model Selection and Training:** Experiment with different regression models (linear regression, random forest, etc.) and evaluate performance metrics.
5. **Feature Importance Analysis:** Identify the most influential factors affecting car prices.
6. **Depreciation trend Calculation and Visualization:** Calculate and visualize depreciation trends for different brands and models.
7. **Report Writing and Presentation:** Prepare a comprehensive report summarizing findings, insights, and recommendations.