Car prediction

Members: Gregor Uustalu, Kristo Kontse
Github: https://github.com/kristokontse/Car-Prediction

Task 2. Business understanding (0.5 point)

Identifying your business goals

Background

The current used car market often suffers from significant price inaccuracies. This causes problems for or budget-conscious consumers and those who lack car expertise and may fall for the overpriced cars. Our project aims to address this issue by enhancing the efficiency of the car valuation process.

Business goals

Our main goal is to create a model that can accurately predict the price of a car based on its features. This helps individuals who don't know much about cars check whether a listed price makes sense. It also helps businesses offer a fair price to customers.

Business success criteria

Business success is achieved when the model produces price estimates that are close to real market values and makes it easier for inexperienced buyers or sellers to understand car prices. Additional signs of success include reduced need for expert appraisal, increased customer confidence, and more transparent pricing.

Assessing your situation

Inventory of resources

Primary Data Source: Used car database sourced from Kaggle which we cleaned and removed incorrect values.

Key Features Utilized: model, year, price, transmission, mileage. fuelType, engineSize, manufacturer.

Programs and language: Jupyter Notebook, Python

Requirements, assumptions, and constraints

The model requires a large and diverse dataset to make accurate predictions for different car types. It is assumed that the listed prices reflect real market value. Constraints include missing or inconsistent data and the fact that car prices may change over time.

Risks and contingencies

Primary risk was data quality - incomplete entries, incorrect values and extreme outliers. We mitigated this risk through a data cleaning process, removing inconsistent data and handling missing values and Nan-s.

The second significant risk is model overfitting - Given our large dataset with many similar car listings and patterns, there is a high chance that models will memorize these. To mitigate overfitting we are data splitting - we use strict train, validation and test split to minimize the risk of overfitting. Also we will apply cross-validation to get a robust estimate of model performance with right parameters.

Terminology
In this project “features” are the car attributes used to make predictions. The “target” is the car price we want to predict. The system is a “regression model” meaning it predicts numeric values. Performance is measured with metrics such as MAE and RMSE, but the final result must remain understandable and helpful to the end user.

Costs and benefits
The only costs our project involves are the time and effort needed to clean and prepare the data, train the model, and evaluate its performance. Our project benefits people who are not familiar with cars so that they can easily check whether the car price is fair.

Defining your data-mining goals

Data-mining goals
Building an accurate regression model to predict car prices from input features. Identifying the most influential factors affecting the pricing. Reduce prediction errors compared to existing estimation methods.

Data-mining success criteria
Our aim is to reach high accuracy on unseen data and predict stable estimates on different cars. Results should align with realistic market prices.

Task 3. Data understanding (1 points)

Gathering data

Outline data requirements
The project requires a dataset that contains detailed attributes of used cars along with their selling prices. The data must include: Car price, car specifications such as brand and model, year of manufacture, mileage, fuel type, engine size (liters) and transmission. Additionally the dataset should contain at least over 20 000 car sale records and car listings should originate from Europe to ensure correct predictions.

Verify data availability
The data was obtained from a public Kaggle dataset containing over 90 000 car sale records. All required features are present in the dataset and fields are generally consistent. Each feature was checked and reviewed, and small adjustments were made to prepare the data for model development. In addition the dataset includes a wide variety of different cars listings which ensures sufficient diversity. The dataset also spans over 25 years of car sales to allow capturing trends and preferences over time.

Define selection criteria
We selected features that have a known impact on the car price. These include both numerical attributes, such as year of manufacture, mileage, and engine size, and categorical attributes, such as make, model, and fuel type. Our dataset did include a tax-rate feature, which we deleted because it is not necessary for predicting car prices and is different in every country.

Describing data
Our analysis shows that our data includes most well-known car brands such as Škoda, Audi, BMW, Mercedes-Benz, and other popular manufacturers. The dataset also has a wide distribution of vehicle prices, ages, and mileages, with newer cars generally having higher prices and lower mileages and older cars having lower prices and higher mileages.

Exploring data
Initial investigation of the data revealed that car prices are strongly influenced by the manufacturer and models. Special models seem to cost more than the standards. Older cars tend to have higher mileage compared to the newer vehicles. Newer cars mostly use automatic transmission, which also contributes to higher prices compared to manual cars. Price and year are strongly positively correlated, meaning that newer cars generally have higher prices. Similar pattern is observed with the pricing and engine size, bigger engines cost way more than the less powerful vehicles. Some of the anomalies were also identified. Certain older cars with rare models had unusually high prices. These outliers will be addressed during the data correction process.

Verifying data quality
Data quality checks showed that there were no missing values in the dataset and numerical ranges are consistent. The dataset also had some extremes, such as very expensive cars and cars with high mileages. To ensure the regression model can predict a wide range of real-world scenarios without losing its accuracy, these values will be kept in. Additionally, categorical variables, such as make, model, fuel type, and transmission, were checked for consistency and standardized where necessary.

Task 4. Planning your project (0.25 points)

1. Data Collection and Understanding
Description: Collecting and exploring car sale databases from Kaggle, ensuring all the necessary features for prediction are included in the dataset.

Tasks: Review available databases and select the one that contains all relevant key attributes.

Team Contribution: Both team members will participate in searching and selecting the most suitable database for the project. Estimated time: 2 hours per member.

Methods/Tools: Google, Kaggle, Data repositories

2. Data Cleaning and Preparation

Description: Review the data to ensure it is clean and consistent

Tasks: Standardize categorical variables, normalize numerical values, remove irrelevant features, and retain extreme but valid values

Team Contribution:
Gregor: 7 hours
Kristo: 3 hours

Methods/Tools: Jupyter Notebook, Python

3. Feature Engineering

Description: Identify and select most relevant features for the prediction model (also choosing the right model) and analyze correlations between them

Team Contribution:
Kristo: 5 hours
Gregor: 2 hours

Methods/Tools: Python, Jupyter Notebook, Panda, NumPy, Scikit-Learn

Comments: Effective feature engineering is critical to improve the model accuracy and prediction performance,

4. Model Building and Training

Description: Train a regression model to predict car prices based on the cleaned dataset.

Tasks: Split the dataset into training and test sets, select regression model, and test different models to select the best performing

Team Contribution:
Gregor: 10 hours
Kristo: 10 hours

Methods/Tools: Jupyter Notebook, Python

5. Model Visualization

Description: Visualize the model through diagrams to show its performance

Tasks: Create plots comparing predicted vs actual prices, generate feature importance charts to show which features influenced the price the most, and plot error distributions to show where the model struggles

Team Contribution:
Gregor: 5 hours
Kristo: 5 hours

Methods/Tools: Jupyter Notebook, Python with libraries

6. Poster and

Description: Prepare a clear and informative visual poster for summarizing the project and get ready for the presentation.

Tasks: Create visually appealing poster that highlight important data, diagrams and results. Prepare a short overview of the project and the plan to present the project.

Team Contribution:
Kristo: 4 hours
Gregor: 4 hours.

Methods/Tools: Design software for creating and customizing the poster (Canva, Photoshop, Figma)

Comments: The poster should highlight key findings and model performance in a visually appealing way, presenting conclusions from the results of the project.