Data science project report

Artur Tamm, Martin Kõnnussaar
Repository link: https://github.com/libakoer/Used_Car_Prices

Business understanding:

Business Goals:

Business Goals:

- Develop a predictive model for estimating used car prices based on various factors such as car make, model, year, mileage, and other attributes. Learn from the development process and apply it later in other things.

Business Success Criteria:

- The model's accuracy will be deemed successful if its predicted price is within 15% of the actual price.

Background:

- The dataset includes a wide variety of features related to used cars, such as their age, mileage, make, model, price, location, etc. This information is intended to help users looking to buy used cars by providing an estimate of the price based on these factors.

Situation

Inventory of resources - two data sets, craigslist data mining function, two people working on data mining, at least 3 PCs of which at least one is fairly computationally potent, python and libraries.
Requirements, assumptions, and constraints - the project must be complete and visualized on a poster by December 13.
Risks and contingencies - Risk: not putting in enough time, procrastinating. Contingency: working on the project on a regular basis, communicating within the team and regularly checking on each other.
Terminology - Target Variable: The variable that the model aims to predict, in this case, the price of used cars.Features: The input variables used for predictions, including make, model, year, mileage, condition, location, etc.Outliers: Data points that differ significantly from the rest of the

data and could affect model accuracy.Imputation: Filling in missing data points, typically with the mean, median, or most frequent value.

Costs and benefits - costs: time, sanity, computation hours. Benefit: way to make americans happy with our accurate predictions on american used car markets

Data-mining goals

Data-mining goals - model that is sufficiently complex and accurate
Data-mining success criteria - knowing more about data mining etc than we did before, have enough information to make good poster

Data understanding

Requirements: Car price, location, condition, year, manufacturer, model, fuel, odometer, transmission
Availability:The dataset contains detailed listings of used cars from Craigslist. It includes multiple features about the cars such as price, make, model, mileage, year, and other relevant attributes. By inspecting the dataset on Kaggle, we can confirm the availability of all the features required for the analysis.
Criteria:Target Variable: price – the dependent variable for the prediction model.
Features: make, model, year, mileage, condition, location, car type, and possibly color, fuel, and transmission
Describing data:
Price: The target variable (continuous), which is the price at which the car is being sold.
Model: Categorical variables representing the brand and model of the car. These can have high cardinality (many unique values), and encoding methods such as one-hot encoding or label encoding may be used.
Year: A numerical feature representing the year of manufacture of the car. The age of the car can influence its depreciation and thus its price.
Mileage: A continuous variable indicating how many miles the car has been driven. This feature is often inversely correlated with the price of a car.
Condition: A categorical variable (e.g., "new", "excellent", "good", "fair", "salvage") that represents the overall state of the car. This can significantly impact the price.
Location: A categorical variable indicating the geographical location of the car. Prices can vary based on location due to market demand, taxes, and regional economic conditions.
Transmission: Cars with different transmissions will have a different price.

Exploring data:

Found issues: Missing mileage, condition, model, price. The missing inputs may impact the results.There are also inconsistent categorical values.

Data preparation ideas:Condition, if missing then the most frequent value will be used. If the price is missing or 0, then remove that row. Fix all Inconsistent categorical values.

Data quality: Missing Values: A preliminary check for missing values in key features such as price, year, and mileage was conducted. Missing values can hinder model performance, and the treatment of missing data (imputation or removal) will be decided based on their frequency and importance.

Data Types and Consistency:

Ensure that numerical columns (like price, mileage, and year) are correctly formatted as numerical data.
Categorical columns like make, model, and condition must be checked for consistency in naming (e.g., "good" vs. "Good") to avoid unnecessary variations.

Planning

1.Data Understanding: Both of us will preform a in-depth exploration of the dataset, do determine what we can use, what we will use and if there might be any problems coming our way. 5 hours minimum.
2. Data cleaning and preprocessing: In this task, we will clean and preprocess the data by handling missing values and dealing with categorical variables. This phase prepares the data for model training. Artur - 10 hours.
3.Model selection and training: The team will select appropriate machine learning models for regression tasks, train them on the cleaned data, and tune the hyperparameters. Martin 10-hours.
4.Model Evaluation and Improvement:Evaluate the performance of the models using various metrics.5-hours Artur, 5-hours Martin
5.Deployment and poster: Deplyoying the model and making the poster for poster session. Both 10 hours.
Tools: Python, Pandas, Numpy, matplotlib, seaborn, randomforest, scikit-learn, jupyter notebook, GridSearchCV, Word, Github.