Group Arizona:
Abibou Mbodji 1931937
Casey Bernal 1625163
Patrick Raogninirina, 1916029

Project Idea: Used Car Price Estimator

Objective:

In this project, we are trying to predict the price of a used car by utilizing model, year, manufacturer, condition, and cylinders as provided by the dataset. The object of the project is to use logistic regression to create a prediction model for different cars.

Questions:

- What attribute has the most impact on a car's price?
- Which attribute holds the least impact on a car's price?
- What formula can be used to predict a car's price?
- Does a car's popularity help the price projection?
- Is there a noticeable price drop after certain years from make?
- What cars hold more value than others?
- Do certain types of vehicles hold more value over others?
- What will the price distributions per attribute look like?
- What factors will contribute to outliers?

Dataset:

The dataset can be found here, it includes a column for the model, the year, the manufacturer, the condition and the cylinders of each car. This dataset was provided by the compilation of craigslist listings that is updated every few months. Most of the data within this dataset contains all the information needed to predict the price of that car in a future listing.

Algorithm:

Some data cleaning and preprocessing are the first task to be completed. The missing data needs to be filled or dropped. We might also perform some dimensionality reduction on the dataset as some attributes are irrelevant to the analysis (cannot be used or has no value/ impact on the model).

Our prediction will be classified in an interval of prices. We will use Random Forest as a tool to both predict the price of a car and to analyze the influence of each attribute. It will also work well with our dataset as it can handle categorical features well as well as large dimensions. We will also use K- Nearest Neighbor (KNN) as a predictor to further highlight the independence between features and also to handle the possibility of a zero-probability of a class. An extra step to encode the categorical features might be needed here to use the KNN model. Finally we will use a Neural Network model for its robustness and learning ability.

We will compare the models based on accuracy and their performance such as cost and robustness.