Final Project Proposal

March 2024

Group Members: Sunny Sun (xs275), Sanjana Vakacherla (sav65)
ORIE 4741: Learning With Big Messy Data
https://github.com/sunny525s/ORIE-4741-Final-Project

Dataset from Kaggle.com:
https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction

Most individuals in modern society use cars as a form of transportation to jobs, schools, and shopping making driving one of the most accessible means of transportation. The used car market is a significant segment of the automotive industry, offering more accessible vehicle options for many individuals. Accurate price prediction for used cars can benefit both sellers and buyers by providing fair market value assessments. This project aims to develop a predictive model that determines a used car's market value based on various features, such as make, model, year, mileage, condition, and location sold. This analysis will not only help in understanding the depreciating value of cars but also assist in making informed decisions in the used car market.

For this project, we consider a dataset of used cars from different parts of India and try to predict the price of each car based on 14 explanatory variables. We obtained the dataset from "Used Car Price Prediction" on Kaggle and the training dataset is comprised of 6019 entries. A test dataset with 13 explanatory variables and 1234 entries is also provided for testing the accuracy of our model. By using the 14 explanatory variables in the training dataset, we will build statistical models with feature information such as the brand and model of the car, the location in which the car is being sold, the year/edition of the model, total kilometers driven in the car by previous owners, fuel type, transmission type, ownership, standard mileage offered by the car, and engine. These feature vectors are generalizable but specific enough such that given any used car in India, the model can come up with a general picture of what the price should be, justifying the price that a used car is sold or bought. With this dataset, it's possible to construct a model to accurately predict the price of used cars given some feature information. Additionally, we hope to use relevant features of this dataset to 1) make accurate used car price predictions based on their most relevant features and 2) create an analysis to determine which features tend to affect the end price of used cars the most.

To answer these questions, we'll use a combination of data analysis and modeling methods to derive valuable insights. Data analysis and visualization methods can help us determine which features best correlate with higher prices of used cars, while various models (linear regression, SVM) can make accurate predictions of the prices themselves. Additionally, we'll clean and transform parts of the dataset through feature engineering to make sure to clear out any null values in the dataset to avoid errors later on in our analysis.

The approach that we will use to answer this question will consist of using different machine learning techniques such as; linear regression, feature engineering, and gradient descent, in order to create models that can help us better predict which factors contribute the most to the selling price of a used car. We can use linear regression in order to identify variables that have a linear relationship. This will help us to establish the basic driving factors behind the resale value of a used car. In the care resale business, there are bound to be outliers - cars that have sold for an exceptionally low or high price - and we can use Support Vector Margins (SVM) in conjuction with linear regression (or other techniques) in order to create models that are more robust to these outliers. SVM will allow us to create predicts while minimizing the impact of outliers on our predictions. Feature engineering can also assist us in identifying useful relationships by allowing us to manipulate our input in ways such as transforming non-numeric inputs into numerical form. For example, we can use feature engineering to create a numeric representation of the make and model of a car. This would make using this information as input for techniques that require numeric input much easier, and enable us to find relationships that would've been difficult to identify otherwise. Finally, gradient descent and regularization could also prove to be useful techniques to use while trying to find the biggest factors in the resale price of a car. Regularization can assist in creating a model that does not overfit to the data, providing a smoother prediction curve. Gradient descent can also be used with multiple loss functions (Mean Squared Error (MSE), L1 Loss, Hinge Loss, etc.), which could be useful if we want to minimize multiple different loss functions for our data.

Being able to predict the selling price of a used car (given additional information about the car) can prove to be very helpful to car retailers, anyone looking to resell a car, and the automotive industry. Car dealerships can use the given predictions in order to make informed decisions about the listing price of used cars in order to maximize profit and number of sales, and first time sellers can use the model to get an idea of what price to expect for their car. The size of the automotive industry in India and the sheer amount of different sellers, dealerships, and independent sellers make this a valuable project to take on. By being able to take in a large variety of input features and output a price, this model will also be usable by different people with different information. In addition, predicting the resale price of used cars can also prove to be useful for potential buyers. They will be able to get an idea of what they should be expecting to pay for a car and seek out the best deal, and can avoid paying an unnecessary additional amount for a car that should've sold for much less. The applications of this project also include potential expansion to cover all automotive vehicles, not just cars. In India, vehicles such a auto rickshaws, motorcycles, and scooters are just as prevalent if not more prevalent than cars. The success of this model could mean expanding it to include these vehicles as well, providing resale information to more buyers and sellers.