Machine Learning for Used Car Price Prediction

COMP 562 Final Project

Peter Yao, Xuan Bai, James He, Ziqian Zhao

May 8, 2022

1 Introduction

The used car market is a significant sector of the US economy, with 40.9 million cars sold and a value of $196 billion in 2021 [2][3]. A wide set of features determine the price of a used car, including age, mileage, and model [7]. Several models have used used in the past to predict the value of used cars based on these features, including machine learning tools like XGBoost, random forest, and lasso regression [3][4]. We evaluated the effectiveness of several machine learning models on predicting the value of used cars Three regression models were selected and trained on a dataset of used car transactions on eBay. The accuracy and speed of the models was then tested, and one of them was selected to be used as the basis of a fourth ensembling model that we developed. The accuracy, speed, and feature importance of the four models was then compared and analyzed.

2 Dataset and Cleaning

2.1 Dataset

Our dataset was a tabular dataset sourced from Kaggle and consists of approximately 160,000 sales records of US and Canadian used cars on eBay over a 20 month period from 2019 to 2020. The dataset contains 13 columns which list ID (a unique value assigned to each transaction), body type (sedan, SUV, etc), number of cylinders, drive type (RWD, 4WD, etc), price sold, year (the year the car was manufactured), zipcode, mileage, make, model, year sold, trim, and engine.

2.2 Data Cleaning

Not all features in our dataset were useful, and many entries contained null values or errors. 4 features deemed largely irrelevant or which had too many errors and null values were dropped from the dataset

We excluded rows with Canadian zipcodes from the dataset, and as several zipcodes were missing the last 2 digits, we dropped the last 2 digits from all zipcodes. The first three digits of a zipcode, while not as specific as the full zipcode, are still able to provide meaningful geographic information about the location of the car's sale.

We considered dropping the feature for number of cylinders as well, as many gas-powered cars were listed as having 0 cylinders in what were likely errors, but this would cause all electric vehicles to be excluded from the dataset as well. Instead, cars with 0 cylinders were dropped from the dataset unless they were Teslas, as the number of non-Tesla electric vehicles in the dataset was very small.

Rows with outlier values in any of the columns (these often contained typos) were dropped, as well as any remaining rows with null values. In the end, about 60,000 of the original 160,000 entries remained.

2.3 Data Visualization

We investigate

3 Models and Results

4 Discussion