Using Data Mining to Build a Car Pricing Model Final Project Report

OVERVIEW

This notebook explores a dataset from kaggle containing data on 426,000 used cars listed for sale. The goal is to understand what factors make a car more or less expensive for my client, a used car dealership. The results of the analysis are presented as recommendations as to what consumers value in a used car and a pricing model froecasting the sales price of new listings. The analysis was conducted using the CRISP-DM framework.

1. Business Understanding

1.1 Objectives

Primary objective

The client objective is to identify the primary drivers of used car value.

Business Questions

When a new vehicle arrives, it is assessed and then listed for sale. What attributes are most valued by customers? How attractive is the vehicle to customers? What should the initial price of the vehicle be?

Success criteria

This project will first identify the top 3-5 vehicle characteristics that determine the sales price. A pricing model tool will allow the client to set an attractive initial price for new vehicle listings.

1.2 Situation

Resources

We have secured a dataset consisting of nearly 500,000 used car listings on Craigslist sites across the U.S. The project is led by consultant with expertise in data mining and modeling. The analysis will be performed using a Python coding in a Jupyter notebook running in Google CoLab.

Assumptions and Constraints

Since we are using a national dataset, we are forced to assume that used car preferences are consistent across regions. We do not have sufficient data to identify differences in our our particular local market, for example, whether customers value specific vehicle attributes more or less highly than the national average.

Also, since the data is based on listings and not sales, we do not know the ultimate sales prices of the vehicles. Items listed on Craigslist often sell for less than the listed price, once the parties negotiate.

Risks and contingencies

We do not anticipate significant risks or delays to successful project delivery.

1.3 Data Mining Goals

The data mining goals are to a) identify the most valuable attributes of used cars, and b) predict the sales price of a vehicle.

Success Criteria

A pricing model which achieves a minimum of 80% accuracy.

1.4 Project Plan

Steps

1. Data Understanding. The data will be explored and assessed for quality, including the distributions of each attribute and identification of missing values.

2. Data Preparation. Attributes will be selected for inclusion in models, with imputation of missing values as needed, and construction of new attributes.

3. Modeling. Appropriate modeling techniques will be applied, including testing fit, computing accuracy metrics, and tuning model parameters to improve both.

4. Evaluation. Data mining and modeling results will be reviewed and evaluated against the business and data mining goals and success criteria.

5. Deployment. A final report and plan for model deployment, monitoring and maintenance will be provided.

Tools and Techniques

Linear regression on the target feature "price" will be the primary modeling technique used in this analysis.

2. Data Understanding

2.1 Data Description

The dataset was sourced from Kaggle but originally scraped from Craigslist. It has 426,000 unique vehicle listings with 18 attributes for each listing, including location and vehicle attributes, year of sale and sales price.

2.2 Data Exploration

Link to charts

2.3 Data Cleaning

The initial look at the data has revealed a couple of issues that need to be cleaned up before applying any filters.

1. Replace all values of "0" with null values, to prevent skewing the descriptive statistics, particularly for in Price, Odometer, and Year.
2. Although there are no duplicated records, there are many duplicate listings for the same vehicle VIN. For example, users listed their cars in multiple regions or relisted on a later date for a lower price. Duplicate VINs need to be removed. Listings will first be sorted by price such that only the listing with the lowest price is retained, reflecting the lack of customer interest at higher prices.

2.4 Data Filtering

After removing duplicate VINs, the dataset consists of 279,288 unique vehicles. It was then filtered to drop unrealistic listings and according to the client's policies:

- Only clean titles and no salvage vehicles
- No buses
- No vehicles over 25 years old
- No vehicles listed with less than 1,000 or over 150,000 miles
- No vehicles priced under 1,000 or over 50,000 dollars

2.5 Data Quality

The final cleaned dataset has 149,861 vehicle listings. Now we will look at the descriptive statistics, with visualizations, and make some observations.

Summary of Observations

All or nearly all listings included these attributes: Price, Year, Manufacturer, Model, Fuel, Odometer, Transmission.

- The average listing price was 14,900 dollars
- The average odometer mileage was 89,900 miles
- The average vehicle year was 2013
- The top 3 most popular manufacturers were: Ford, Chevrolet, and Toyota
- 90% of listings were for gas fueled cars
- 88% of listings had an automatic transmission

There was only partial data for these attributes: Condition, Cylinders, Drive, Size, Type, Paint Color. Among listings that included these attributes, the majority were for:

- Engine size - large 6-cylinder (23%) or 8-cylinder (13%) engines, rather than 4-cylinder (25%) engines
- Vehicle condition - Excellent (30%) or Good (19%), followed by Like New (8%)
- Vehicle type - Sedans (23%), or SUVs (20%), followed by Pickups and Trucks (12%)
- Vehicle size - Full size (16%), followed by Mid-size (12%) and Compact (6%)
- Drive – 4WD (29%), Forward (30%) and Rear (12%)
- Paint Color – White (17%), Black (15%), and Silver (11%)

The three numeric variables are all correlated. Price is negatively correlated with age and odometer, and age and odometer are positively correlated with each other.

2.6. Data Preparation

The dataset was prepared for modeling by dropping string variables which have too much or too little heterogenity, or do not contribute to customer appeal: VIN, region, model, type, title_status, clean_title, paint_color, state, transmission, fuel. Further steps were incorporated into the modeling process.

3. Modeling

Linear regression was the primary modeling technique. Various strategies for feature inclusion will be tested. Models will be evaluated using mean squared error and by goodness of fit (R squared) metrics. Model parameters were tuned to improve both.

3.1 Model test design

- The data was split into target and independent features, using the log of the target feature, “price”.
- It was then split into training and test sets, with a ratio of 30% for testing.

- A baseline model was then calculated using the mean values of the training and test sets, and evaluated using the mean squared error between predicted and actual values: Train MSE was .519 and Test MSSE was .524

3.2 Feature preparation and transformation

The clean dataset with 7 features was transformed for modeling using a preprocessing step. As a result, the dataset used for modeling had 45 features and no missing values.

- Feature Year was converted to Age
- All the numeric features skewed (non-Gaussian) distributions and were therefore normalized (re-scaled)
- Missing values were imputed using “most frequent” strategy
- Categorical features in string format were encoded as numerical dummy features for ingestion in the model
- A new ordinal feature was created from the “condition” feature representing increasingly better condition with increasingly higher numbers

3.3 Models

Several variations of linear regression were performed:

- Model 1. OLS linear regression (Oridinary Least Squares)
- Model 2. Feature selection in preprocessor, then OLS linear regression
- The most statistically significant features were selecting using the Select Percentile strategy, 10th percentile
- Model 3. Ridge linear regression with grid search for best alpha

3.4 Summary of Model Results

The models were evaluated by comparing the mean squared error between the predicted and the actual values in the training set and the test set.

Both Models 1 and 3 used the full dataset of 45 features and had very similar results. The alpha value for the Ridge regression in Model 3 was .001, indicating that very little weighting was applied to the model coefficients. It could not improve much, if at all, upon the OLS, or standard linear regression, used in Model 1. Both of these models had very good performance, with a mean squared error (MSE) of .189 and R-squared of .637 for the test set. The low MSE value of is indicates that the model’s errors were close to zero. The R-squared indicates the model can explain about 64% of the variability in price, with the remainder due to factors not included in the model.

Model 2 used a selection technique to evaluate each of the 45 features and select only the most statistically significant ones for inclusion in a linear regression model. A threshold of the tenth percentile of significance was used, resulting in the selection of 7 features. The evaluation metrics for this model were not as good as the others. However, the R-squared value of .531 on the test set revealed that these 7 features explained more than half of the sales price.

Highly explanatory features contributing to increased price were pickup and 8 cylinder engine, and those contributing to a lower price were front wheel drive and vehicle age. The features four wheel drive, condition and sedan were also selected, and all had a weak negative impact on price in this model. Since we expect four wheel drive and better condition to have a positive impact on price, that indicates there is some issue with the data, most likely because both of these variables had a significant proportion of nulls.

All the coefficient values in Models 1 and 3 were of the expected sign. These models also identified pickups as strongly influencing higher prices and vehicle age strongly influencing lower prices. Other factors strongly associated with higher prices were 12 cylinder engines, luxury manufacturers, and Ram, GMC, Jeep and Toyota makes. In general prices increased with engine size. A weaker but positive association on price was due to four wheel drive, convertible type, better condition, and Ford, Honda, Chevrolet makes.

Factors strongly associated with lower prices were vehicle age, odometer mileage, hatchback type and Mitsubishi, Saturn, Mercury, Fiat makes. Age had a stronger influence on lower prices than mileage, indicating that customers value a newer car over higher mileage. Factors weakly associated with lower prices were forward sedan or wagon types, 4 or 6 cylinder engines, forward drive, and Chysler, Nissan, Hyundai, and Kia makes.

Summary of Model Evaluation Metrics

	Mean Squared Error	R-squared
Baseline	Train MSE: 0.519 Test MSE: 0.524	
Model 1	Train MSE: 0.185 Test MSE: 0.189	Train R-Squared: 0.643 Test R-Squared: 0.637
Model 2	Train MSE: 0.241 Test MSE: 0.245	Train R-Squared: 0.534 Test R-Squared: 0.531
Model 3	Train MSE: 0.185 Test MSE: 0.189	Train R-Squared: 0.643 Test R-Squared: 0.637

Summary of Coefficients

	Models 1 & 3	Model 2
++	12 cylinders Luxury Pickup Ram, GMC, Jeep, Toyota	Pickup 8 cylinders
+	10 cylinders Convertible 4WD Ford, Honda, Chevrolet Condition	
-	Wagon, Sedan Chysler, Nissan, Hyundai, Kia 4 or 6 cylinders FWD	4WD Condition Sedan
--	Odometer Hatchback Mitsubishi, Saturn, Mercury, Fiat Age	FWD Age

4. Evaluation

The descriptive statistics and linear regression model estimated from the dataset have resulted in useful insights.

4.1 Assessment of Data Mining Results

The data mining goals are to a) identify the most valuable attributes of used cars, and b) predict the sales price of a vehicle. The first goal has been achieved, as shown in the section above. The second goal has also been achieved, as Model 1 could be used to predict a sales price with 64% accuracy.

The success criteria of a pricing model which achieves a minimum of 80% accuracy was not achieved. One limitation was the dataset itself. Although it seemed robust with nearly 500,000 listings, after data cleaning we were left with only 150,000. Many of those listings were missing important nature such as vehicle condition. The models developed here could be improved with a larger and more complete dataset.

4.2 Review of Business Questions

This analysis have achieved the primary client objective of identifying the primary drivers of used car value. Based on the attributes identified, the client should be able to quickly assess a new vehicle when it arrives and calculate an attractive sales price using the pricing model.

5. Deployment

The final model and all documentation may be found at this link