Will

[Email address]

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

[Document title]

[Document subtitle]

Background and Description of Problem

Kelley Blue Book is a vehicle valuation and automotive research company based in Irvine, California. The company provides market value prices for new and used automobiles of all types. Kelley Blue Book products include Fair Purchase Price and Fair Market Range. These value types are based on actual transactions of what individuals are paying for vehicles and are adjusted accordingly as market conditions change on a weekly basis. For used cars, KBB also provides suggested retail value, trade-in value, private party value, certified pre-owned value, and auction value. As technology becomes more accessible and affordable, new players are popping up in the vehicle valuation industry. Companies like TrueCar and Carvana introduce their own values to compete in the already saturated industry. It is for this reason that Kelley Blue Book must stay competitive by exploring other methods that will produce more accurate and reliable values for the consumer.

Since Kelley Blue Book receives data from several source including dealers, auctions, and other external partners, it would be ideal to train a statistical model to predict the sale prices of vehicles. The model will utilize new car transaction data compiled on a weekly basis from an existing external partner. The output of this model will serve as the Fair Purchase Price for new vehicle purchases.

Data and Description of Features

New vehicle transaction data was gathered from one of Kelley Blue Book’s external partners. The data consists of over 285,000 new cars and trucks sales from December 2016. While not all of the features were used to determine the model, it is important to understand how the data is structured. Let’s take a look at some of these features:

Model Year:

The model year of the vehicle sold. Since the data was for transactions in December 2016, we expected most of the model years to be 2016s and 2017s, and indeed this was the case. Interestingly enough, there were also some model year 2014s and 2015s scattered throughout the data.

Make Name:

The manufacturer of the vehicle sold. Nissan leads all makes with over 32,000 new car transactions in December while Alfa Romeo had only one transaction. Exotic manufacturers such as Lamborghini and Ferrari were excluded in the data because it is unusual to see these vehicle sell for under MSRP. Therefore, it would not be considered Fair Market value.

Drivetrain:

The drivetrain configuration of the vehicle sold. These include four-wheel drive (4WD), all-wheel drive (AWD), front-wheel drive (FWD), Rear-wheel drive (RWD), and two-wheel drive (2WD). Depending on the vehicle, the drivetrain configuration might be the same but have different names. For example, 2WD can be either FWD or RWD. However, 4WD and AWD are different due to the way power is distributed to the wheels.

Mileage:

The mileage of the vehicle sold. Since these are new car transaction, we expect the mileage to be relatively low across the data. While the majority of the vehicles sold had less than 100 miles, there were vehicles sold with several hundred miles.

Days in Inventory:

The number of days the vehicle stood in the dealer’s inventory before being purchased. This ranged from less than 1 day to several hundred days with an average of just over 76 days.

Number of transmission speeds:

The number of transmission speeds the vehicle had.

Transmission type:

The transmission of the vehicle sold, either manual or automatic.

Door count:

The number of doors the vehicle had.

Engine Cylinders:

The number of engine cylinders the vehicle had.

Engine Displacement:

The swept volume of pistons inside the engine cylinders measure in liters.

Engine Type:

These include gas, diesel, hybrid, electric, and flexible fuel.

Sale Price Amount:

The amount the new car was sold for.

Feature Selection and Analysis

We will use a correlation matrix to quantify the linear relationship between each of the features and the target, the sale price amount of the vehicle. This also allows us to visualize features that are highly correlated, also known as multicollinearity. While multicollinearity will not reduce the predictive power or reliability of the model as a whole, we might want to reduce it in order to improve the interpretability of the model. 