Background:

We are a secondhand cars dealer.

Car prices have been changing rapidly in recent years due to the pandemic. We want to know how much cars are worth now so we can set prices for our inventories. Knowing the car price is also important for us because we will need to purchase used cars from others. Knowing the car price gains us advantages in inventory purchasing.

Dataset introduction:

The dataset is from crag list. Its sample size is large enough to reflect the real market condition. It also contains various variables covering different status. Those variables can be converted into dummy variables or into levels. In our model, title status and transmission are converted into levels while fuel, type and drive are converted into dummies.

At the end we can use text mining for description, from which we will see the hot topics.

Visualization:

Our goal is to see price comparisons among different variables. Besides the price, we also want to dig deep into marketing information.

The first two plots are price comparison for different fuel and car types.

On the left side is a pie chart comparing average car prices for different fuel types. The chart shows that diesel cars have the highest average car price, and electricity cars have the second highest.

The box plots on the right side shows that Trucks, pickup and offroad have the highest price. Sedan, wagon and minivan have the lowest. There could be an relation between car size and price. The bigger the size, the higher the price.

The second plot is average price comparison for different manufacturers. From the manufacturer's plot, once we receive a car, we will have rough estimates on price. Besides the luxury car brands, Tesla’s price is way higher than popular choices such as Toyota, Honda, Ford, Volkswagen etc.

From the heat map we can see the average price for different states. The deeper color states are concentrated in the mid-west. Lighter color states are in the mid-east, such as OH & PA. Alaska’s average car price is the highest among all states, that could be due to the large amount usage of diesel trucks.

We can further investigate different cities and compare different variables. The plot on the right is a demonstration if we want to compare average price between transmission. The two maps on the next slide shows drives train comparison between Pittsburgh and Columbus. Summarized up, we can pinpoint customers’ need based on their location.

Model:

We first make the correlation plot. Deeper color shows there is potential relationship existing. By looking through the first row, year, condition, cylinders, size, diesel fuel, truck type, and 4-wheel drive appear to be related.

We then set a linear model, which returns an R-square of 0.60. These numbers are good enough for real world data. And we can use this dataset to make further modeling investigation.

Next, we create a base regression tree. The tree’s OSR is 0.73. Would say this tree is good enough to be used for price prediction.

But we want to go deeper. We created a random forest including 300 trees. We figure out the best MTRY value for random forest is 16 and from the nude purity we see the importance ranking for different variables. Comparing the nude purity and our previous correlation map and linear model, we are glad to see that our previous thoughts are very close. Beside those variables which have already shown a correlation, the odometer appears to be the second important variable.

The random forest’s performance is reflected by looking at the OSR, which is 0.86.

The last model we created is XG boost, this model has a good performance but not as good as the random forest. OSR is 0.83

Finally, we sort out our 4 different models and compare their performance between MAE and R-square.

Text mining:

Remember we said at the beginning. Our dataset includes a section of car descriptions. We do text mining and sort out the words that appear the most.

These words reflect the current market situation, which car type is popular, what do customers want? Based on this, our sales representative can pinpoint customers’ needs accordingly.