Predicting Depreciation of Used Vehicles

Adam Ezzat

Introduction

- What matters to you most when you're looking for a car? Some people are most concerned with how it feels when they pull it off the lot, how it drives, how cool it looks. But some more savvy consumers are interested in a practical aspect - *how well does the car retain its value?* If they want to sell it in a few years, how much of their investment will they be able to recoup?

- Other than direct customers, used car sellers would also be interested to know the answer to this question as it could help them better understand if a car is over or undervalued. In this project, we seek to examine what features predict how much a car will depreciate, and build a model to predict exactly that.

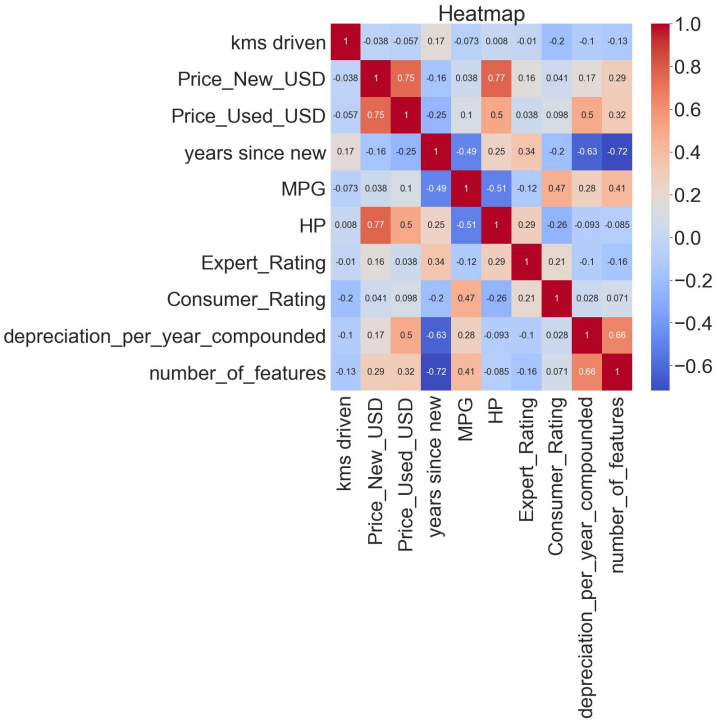
Potential Clients

- The client in this situation would be buyers who are looking for cars that will retain value over time. This product could also be used by companies like Kelly Blue Book looking to help customers do more thorough research into which cars to buy. The recommendations to the clients would be to buy a car with fewer features as they seem to retain their value better over time.

Data

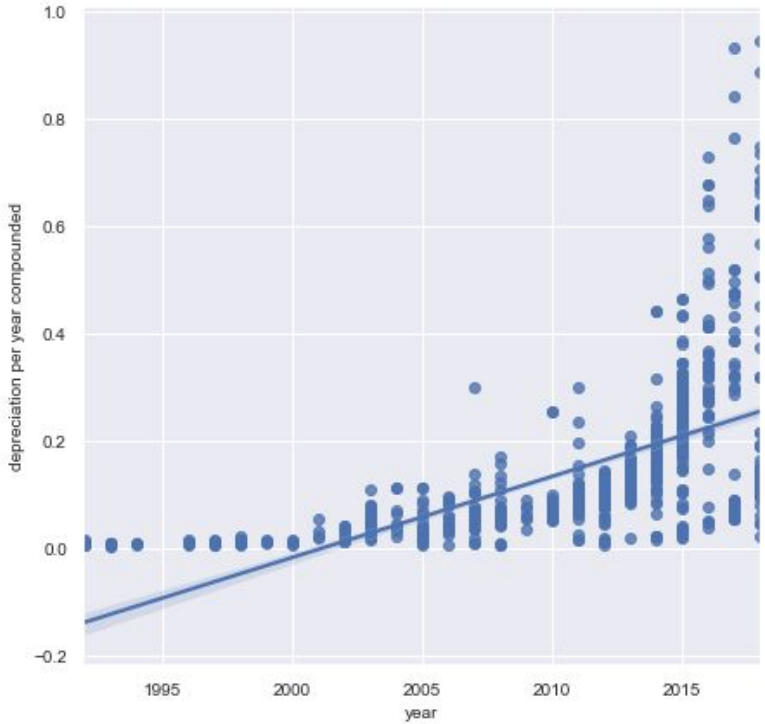
- The data came from Kaggle and it was about used cars that were sold in Pakistan, ultimately trying to figure out which features are responsible for a car retaining its value over time.
- The website that I used to scrape the missing information from was KBB.com(Kelley Blue Book).
  After scraping the data from KBB.com its important to remember that the price that the cars were sold at in the original dataset from Kaggle are in Pakistan Rupee , while the price listed on KBB.com is U.S dollars.

EDA



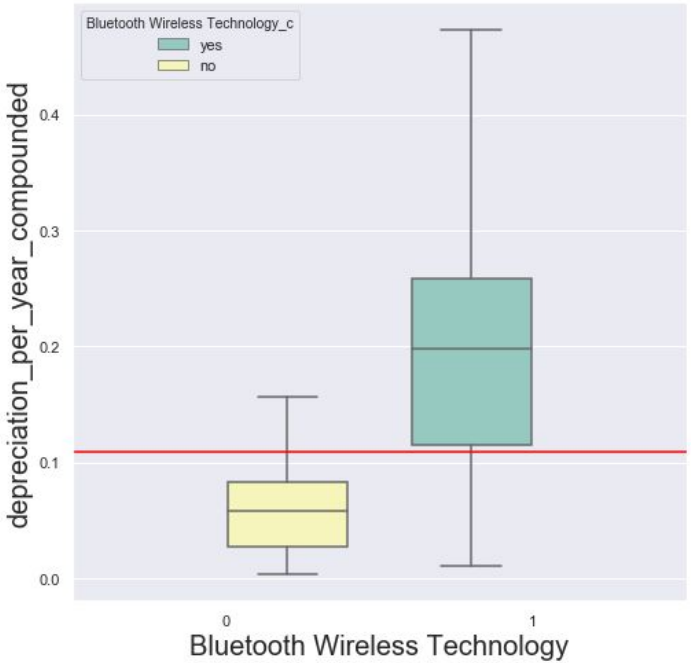
- Correlation Heatmap
  - I used the correlation heatmap to help figure out which features had either positive or negative correlation to depreciation per year compounded so that

Year of the models



- This graph shows that there is a positive correlation to how old the car is to depreciation. The older the car is the less the compounded depreciation will be.

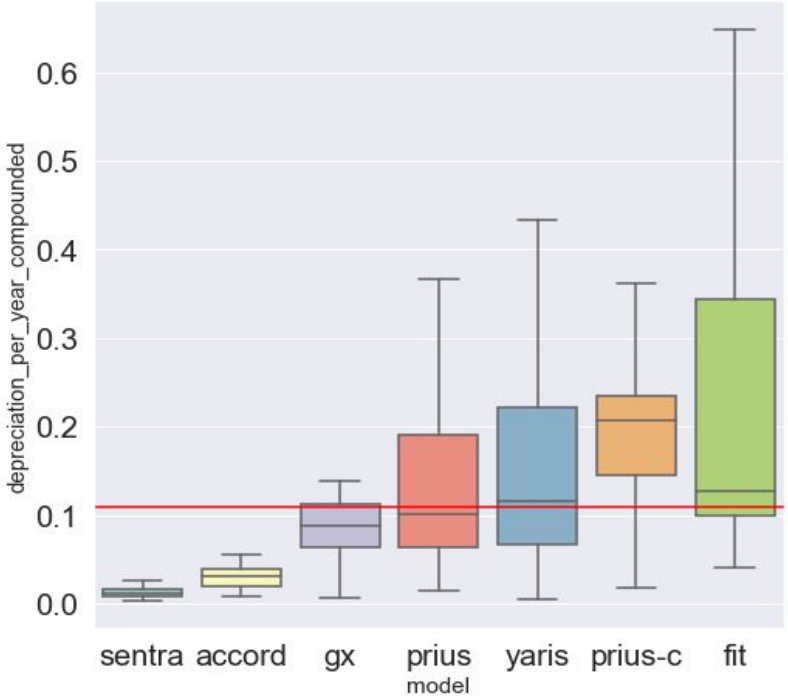
Features



For simplicity sake will cover only one feature as the general outcome is similar for the other features

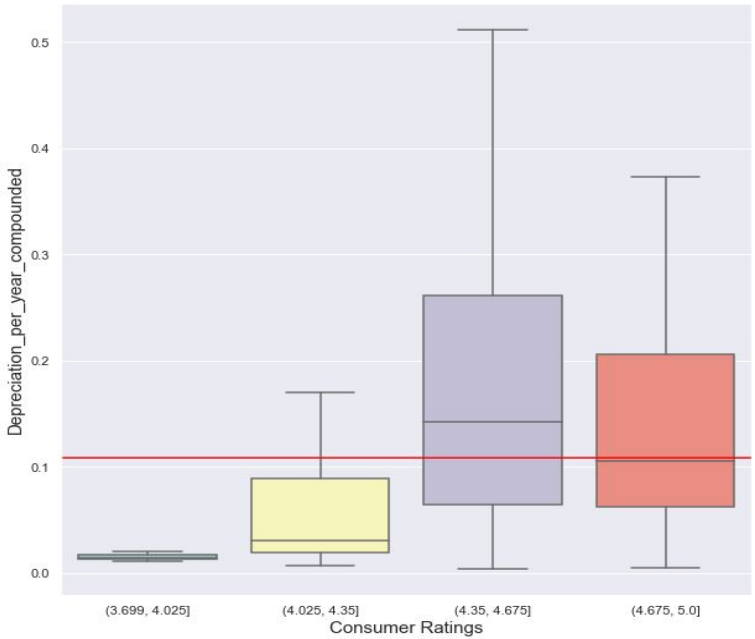
- A few features in particular had a high correlation with depreciation per year. Cars which have bluetooth streaming audio depreciate faster than those without it ($p < .001$) , the reason for this could be the bluetooth in cars will become out of date faster and while the cars that never had one to begin with could always have the latest bluetooth installed into the car.

Models



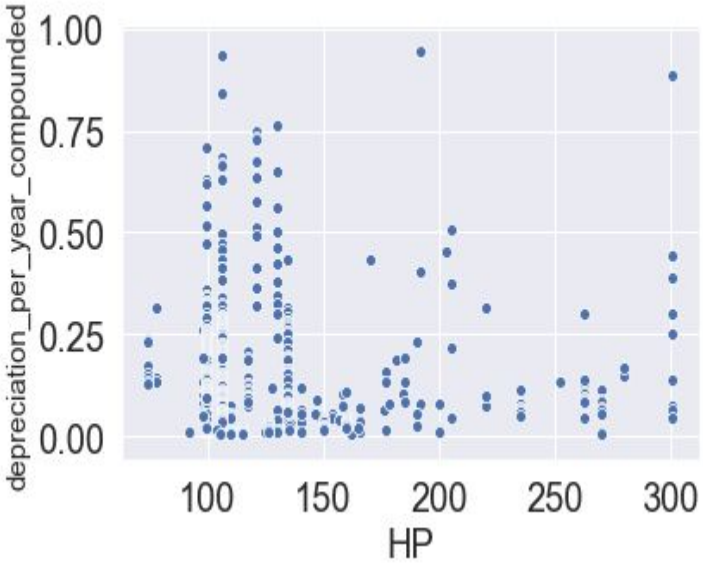
This graph shows the top 7 models in the dataset, from what is shown the Fit was the car which had the greatest depreciation.

Consumer ratings



- Much like the features graph, the consumer ratings was very similar to safety ratings and expert ratings.
  - The general consensus was that consistently faster depreciation with higher ratings even in the highest ratings category.

Story Telling



- The next graph that we will look at is horsepower(HP) in the x-axis and depreciation per year in the y-axis. From what we see here the less HP a car has the more it depreciates per year, this shouldn't be much a surprise as the car, given that a more power car engine would be more desirable to most people even if they have no plans to go racing with the car

Machine Learning

1. The next step was to build a ML model to predict the depreciation per year for a given used car as accurately as possible. I decided to try tree-based methods because they work well when there is a non-linear relationship between the features and the target variable as is the case with a few of our features here (such as “years since new”).
2. I used Gridsearch with cross-validation for each model to optimize the hyperparameters before determining the error against the test set. The results are below:

Machine Learning Part 2

For the results of Random Forest I used the parameters Max depth and n_estimators.

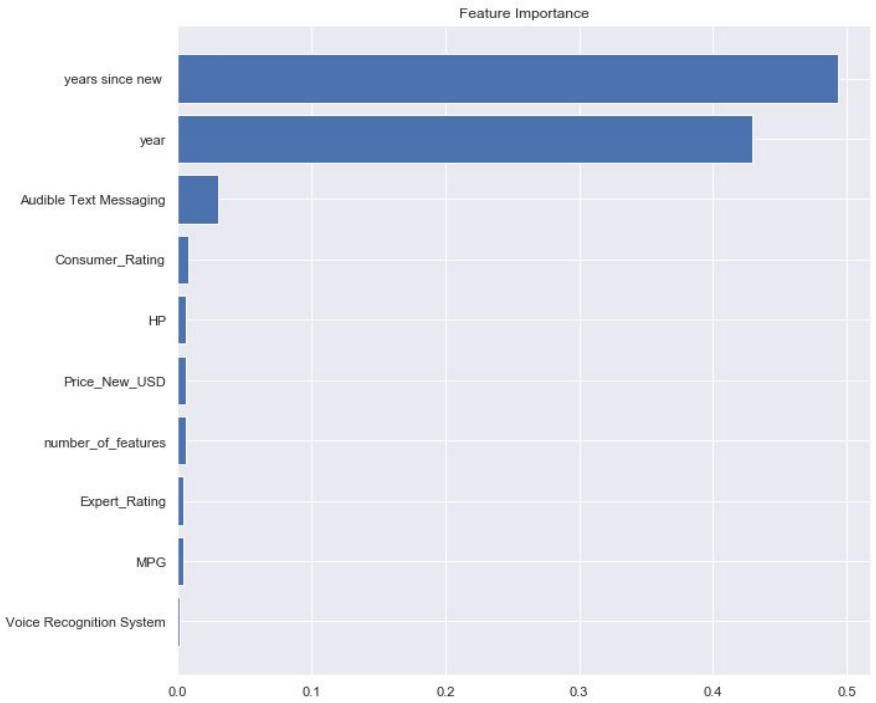
1. For Max Depth I used 20 and n_estimators I used 50
2. The R squared results were 0.693616599681034

For Gradient Boosting Regressor I used the parameters Max depth and n_estimators.

1. For Max Depth I used 3 and n_estimators I used 750
2. The R squared results were 0.6797173263251267

So the best model in this case is the random forest regression model

Feature Importance



If there is one characteristic of why the vehicles retain their value it is the age of the car when sold. Therefore, the largest piece of advice to anyone looking to buy a car with value-retention in mind, is to buy an old used car.

- Ideally, an older used vehicle
- Lacking the newest technology
- Low in price when new
- Rated highly by consumers

Conclusion and Next Steps

- What this model was able to predict was depreciation per year on a new vehicle was 7% error on average. In general I found that the more desirable a car is (features like horsepower and audible text messaging) the faster it would depreciate.
- The next step to give a more complete picture of why cars retain value over time would be to study how many kilometers were driven, the wear and tear of the car would be another useful factor in why a car might retain its value over time.