Used Car Price Prediction

Sujeeth Shetty

BUAN 6347.501

INTRODUCTION

Buying a car is never an easy task, especially when buying a used one. So many different factors go into determining the price of a vehicle that its difficult to accurately predict what one should be paying. Aside from this there are some parts of the country where used car sales are few and far between and the prices are not so desirable.

Through the analysis I want to find interesting trends and correlations between the various features of the sale of the used car. In addition, I want to leverage Ganymede, Python, PySpark and ML for predicting the prices of the used vehicles.

OBJECTIVES

Identify the trends in the sales of cars and car industry in general.

Correlation between price and condition which is the further extended by year, number of cylinders and mileage.

Prediction of a used car's price given its specification .

METHODS

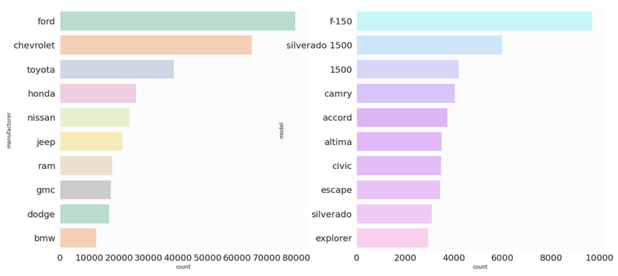
Dataset

The Used Car Dataset from Kaggle is used for this project. The dataset before cleaning has 423857 points with 24 variables. This dataset originally scrapped from Craigslist.org and updated every month. The key features of the dataset are shown below:

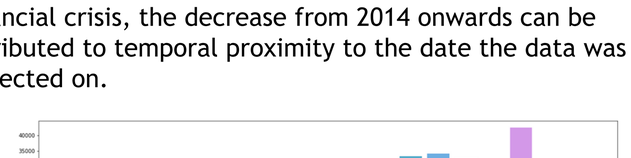
price	region	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission	drive	type	paint_color	state
35990	auburn	2010.0	chevrolet	corvette grand sport	good	8 cylinders	gas	32742.0	clean	other	rwd	other	NaN	al
7500	auburn	2014.0	hyundai	sonata	excellent	4 cylinders	gas	93600.0	clean	automatic	fwd	sedan	NaN	al
4900	auburn	2006.0	bmw	x3 3.0i	good	6 cylinders	gas	87046.0	clean	automatic	NaN	SUV	blue	al

Exploratory Data Analysis

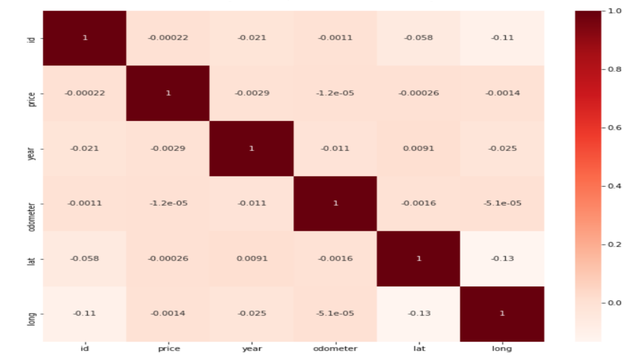
In exploring the dataset, I found that Ford leads the pack with most cars up for sale. This comes as no surprise since Ford's F-250 is among the top models.



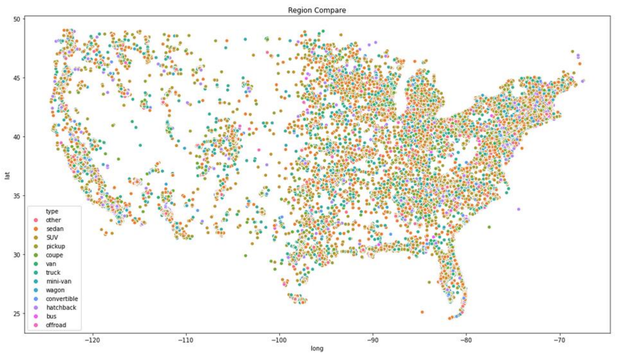
Looking at the trend of number of cars for sale according to the year they were manufactured, there is a decrease in trend in cars available for sales from 2009 which can be due to the 2009 financial crisis, the decrease from 2014 onwards can be attributed to temporal proximity to the date the data was collected on.



The other observations, I found an increase in vehicle median price over the years. Also, a decreasing trend is observed between odometer and price, which is expected.



Its also observed that Sedan type has the greatest number of vehicles. By the comparing the vehicles types across region, I didn't find any significant difference.



Data Engineering

The original dataset had fields which were not directly contributing to the prediction. The first was to remove those columns.

In the next step, created two new features age & region_enc. The age is the difference between the vehicle manufactured year and posted_date. Since state & region were converted to numeric feature called region_enc which is mean price difference across regions.

Performed one-hot encoding on fields with <10 categories. Whereas the fields like Manufacturer and Model had more than 40 & 10000 categories respectively, hence label encoding was applied.

The missing values in fields, condition, cylinders & title_status were imputed by Linear regression.

The odometer value was normalized to reduce the variance. Finally, removed the columns like paint_color which has highest proportion of NaN and cleared rows with Null values.

Modeling

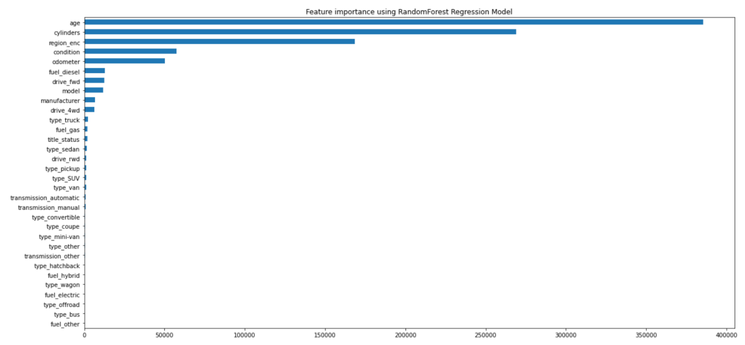
Since the dataset has a greater number of nominal variables, the tree-based model would be a better choice. So, I decided to go with RandomForest as base model and use advanced algorithm like Light GBM & Catboost to get the best predictions.

RESULTS

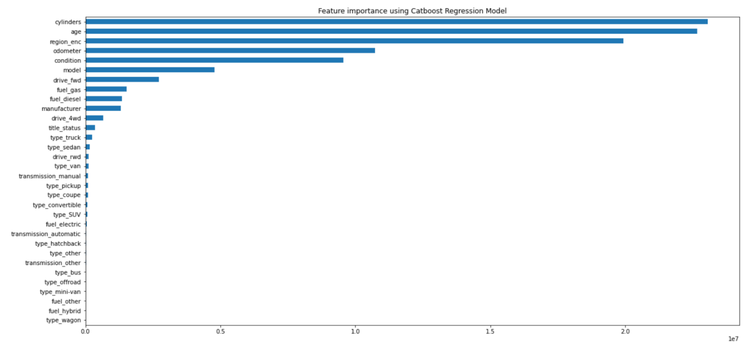
The three models, Random Forest, Light GBM and Catboost were run simultaneously, and the results were recorded. The Catboost algorithm which can handle categorical variable better than any other model performed better and was able to capture 96% of variance. The results of all three models are shown below.

	Random Forest	**Light GBM**	**CatBoost**
R2	95.6%	94.81%	95.93%
MAE	989.21	1446.89	1232.4

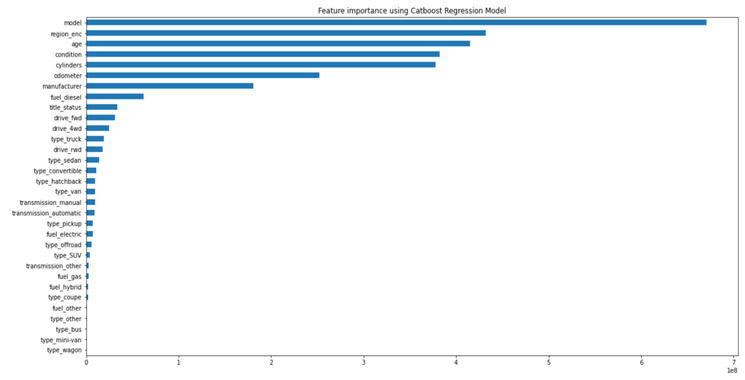
Looking at the feature importance using Random Forest Model, age, cylinders & region_enc has major influence on the prediction.



Even in case of Catboost, cylinders, age & region_enc had more influence.



Compare to other models, Light GBM was the worst performer but it was able to explain 94% of variance. Here, Model feature has more influence on the prediction.



CONCLUSIONS

Looking at the output of model we can conclude that the condition of the vehicle plays important role in vehicle price. Since region_enc is an important feature in all the models we can say that the vehicle price varies across region and it also play important role in prediction.

Even though R2 is >90%, the mean absolute error for all the models are very high too. This highlights that there are more features like taxes, trim, number of owners are out of scope of the dataset but play a significant part in affecting used car price.

This model can be further improved by fine tuning the existing variables like Model. In certain cases, the model values are not accurate and such attributes can be retrieved using NHTSA API's. Also building a sperate model for each Manufacturer helps in reducing dimensions and could also improve the accuracy of the model.

Improving the model further and deploying it as an API will provide an easier experience for those searching for a used car.

REFERENCES

[1] https://ahmadosman.com/BestCarDeal/

[2] https://medium.com/swlh/exploring-and-analyzing-used-car-data-set-2e2bf1f24d52

[3] https://www.kaggle.com/austinreese/craigslist-carstrucks-data

[4] https://www.kaggle.com/anerisavani/eda-and-price-prediction-of-used-vehicles

[5] https://www.kaggle.com/jeongbinpark/craigslist-car-data-analysis

[6] https://catboost.ai/docs/concepts/about.html

[7] https://lightgbm.readthedocs.io/en/latest/