Technical Report: Used Vehicle Price Prediction Using Craigslist Data
Data source (Kaggle): **https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data/data**

Business Problem:

In U.S. buying and selling used cars on craigslist is a common culture but finding it in the right price is a struggle faced by millions. Selling a used car at a high price is difficult and selling it in a low price may lead the seller to lose money. This project has compiled a data from craigslist to build a model that predicts the price of a car, which will help both buyer and seller to make decision based on the existing prices on craigslist.

Data Source and Preprocessing:

To build this model, a public dataset from Kaggle named "*Used Cars Dataset*" was used with listing of cars and trucks from Craigslist. The dataset included records that were missing which was dropped while cleaning it in the process. The model consists of 5,670 records that has about 21 features. One-hot encoding was used to turn the categorical features into numbers. The data for this model was spilt into 80/20 as per the requirement.

Model Selection and their Performance:

The models that were tested are:

1. Simple Linear Regression: Poor prediction (R^2 of 0.10) using only odometer.
2. Multiple Linear Regression: Improved prediction (R^2 of 0.466) on data that was tested.
3. Random Forest: Best result (R^2 of 0.751) with an average error summing up to $4,091

The model prediction was affected by key factors like odometer reading in the car, manufacturing year and age of the car. For example: a car's predicted price by the model was about $18,541.38 with the 95% confidence range from $5,577.32 to $31,505.43.

Key Insights and Recommendations:

Random Forest is best suited for these kinds of pricing prediction models. One of the key important factors, before buying a car that can be observed from the model are the age of the cars and how much is clocked in the odometer. From these observations, a seller can determine and set fair prices which helps the buyer to avoid overpaying.

Limitations and Future Improvements:

There are few limitations in the model which is majorly because of missing information in the dataset. The model doesn't represent the market demand for a car.

To improve the model, updating the data every year will help it remain accurate over the time. The main target is to build a mobile application which will help the model with live data and easier pricing predictions.