**Proposal Document**

**Used Vehicle Buying:**

Improving the Buyer and Seller experience through inference and prediction models

**Group Title:** Group1

**Group Members (In Alphabetical Order):**

Ralph Parlin | Patrick Prioletti | Brian Schramke | Kobi Wiseman

**Research Objective:**

***Problem statement:***Current processes for determining the price of a used vehicle creates pain points for both U.S. vehicle buyers and sellers as they struggle to determine a vehicle’s fair market value. However, the valuation process can be improved through effective modeling.

***Motivation:***The timing for improvement could not be better. COVID-19 has left virtually no corner of the economy untouched, but the automotive industry has been among the hardest hit. However, according to a Wall Street Journal article from July 2020, the used vehicle market has made a tremendous rebound as economies open back up, with sales of used vehicles in the U.S. climbing back faster than that of new vehicles after dropping 38% in April [[1]](#endnote-1). Although there are divergent opinions on the causes, one thing is clear: U.S. consumers are on the hunt for used vehicles.

Buying used vehicles can be challenging. Buyers are unaware of what sellers will take as a reasonable offer for their vehicle. Sellers, to include dealers, are often unaware of what price they call sell their vehicle for. This tension on price is muddied further by factors such as vehicle features, geographic region, time of year the sale is offered and alike. Although in recent years several applications such as CARFAX, Cars.com and alike have assisted, room for improvement remains with respect to informing buyers and sellers on the used vehicle market. We believe our timely analysis will help.

Through the tools and techniques outlined below, this analysis aims to improve the understanding of used vehicle buying through inference gained from regression analysis and to offer forecasting models through linear regression and logistic regression that will help better predict the true value of a used vehicle.

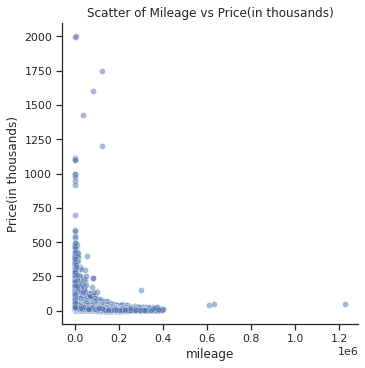
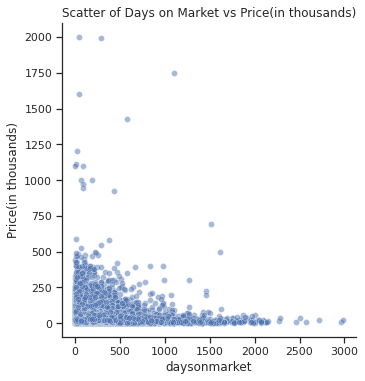
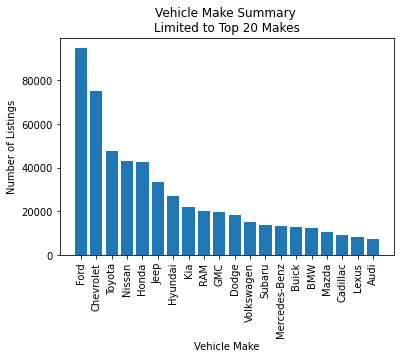
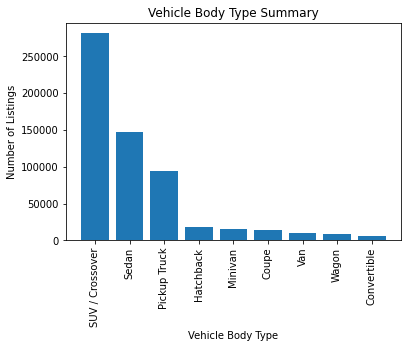
**Data Set Description:**

* Overview / Description
  + The dataset contains real data on U.S. used vehicles from sale data on CarGurus updated on 21 September, 2020
* Number of observations and features
  + 3,000,000 observation across 66 features
* Sample predictors
  + Vehicle attributes [milage, sunroof, body\_type, fuel\_economy]
  + Sale attributes [daysonmarket, is\_certified, listed\_date]
  + Regional attributes [city, dealerzip, lat/long]
* Link to dataset:
  + https://www.kaggle.com/ananaymital/us-used-cars-dataset
* Anything interesting or surprising about the data
  + How many expensive vehicles there are and how many days some of the vehicles are on the market

**Preliminary Data Exploration:**

To reduce the dataset to a manageable size for this research, we randomly selected 600,000 sample observations from the 3 million original records to use as our analysis data frame. After an extensive cleaning process that included the transformation of several features, development of composite features, data type changes, and duplicate row removal, the shape of data frame became 596,000 observations across 58 features. Highlighted below are a small sample of summaries and visualizations that have helped us better understand our data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Summary Statistics** | | | | | |
| **Summary / Feature** | **price** | **city\_fuel\_economy** | **daysonmarket** | **horsepower** | **mileage** |
| **Mean** | 29941.16 | 22.69 | 76.15 | 248.01 | 31121.33 |
| **Stddev** | 19376.94 | 8.84 | 108.99 | 90.51 | 45289.33 |
| **Min** | 165 | 8 | 0 | 67 | 0 |
| **Max** | 2000000 | 127 | 2979 | 808 | 1225238 |

****

**Predictions:**

Using Linear Regression models, we plan to achieve the following:

* Predict the listing price of a vehicle based on a collection of variables. Can be used by sellers to set fair prices and by buyers as benchmarking for value determinations
* Predict the number of days listed based on vehicle configuration, region, and price to help sellers price vehicles to sell in desired timeframe

Using Logistic Regression models, we plan to achieve the following:

* Predict whether a vehicle was previously part of a commercial fleet or a taxi to help buyers avoid heavy use vehicles
* Predict region where to find a particular vehicle configuration to assist buyers in targeting searches

**Inference:**

Using Linear Regression models, we plan to achieve the following:

* Identifying which features/option (sunroof, bucket seats, etc.) on a vehicle provide the greatest/least influence on the price of a vehicle
* Determine the best features for a price prediction model
* Identifying the impact (if any) of any accident that occurred to a vehicle when predicting its price

Using Logistic Regression models to achieve the following:

* Determine the best features for predicting the probability of a vehicle listed on the market for a long time
* Determine how a vehicle’s price (which will be used as a predictor, unlike the linear model) on the probability for a vehicle to be listed on the market for a long time
* Identifying the impact of a vehicle’s accident history has on predicting its probability to stay on the market for a long time

In addition, we also plan to gain inference and inform modeling by experimenting with unsupervised learning through the employment of k-means cluster analysis to see what patterns emerge in the used vehicle market. Specifically, we expect to see clusters form around price, vehicle attributes, vehicle technical properties, or perhaps by geographic region.

**Non-Spark Packages:**

|  |  |
| --- | --- |
| **Non-Spark Package** | **How the package relates/assists our analysis.** |
| pandas | Support work done in spark and help review or visualize data |
| matplotlib | Support work done in spark and to improve basic EDA |
| seaborn | Support work done in spark by improving data visualizations in the report |

1. Wall Street Journal, “During COVID-19 Pandemic, the Used-Car lot is Hot”, Mike Colias, July 2020 https://www.wsj.com/articles/during-covid-19-pandemic-the-used-car-lot-is-hot-11593774001 [↑](#endnote-ref-1)