Predicting Used Car Sales Price using Applied ML

DSCI 631 Group 7: John Johnson, Jacob Linder, Katy Matulay

Dataset

- Source: Kaggle
- 3 million US used cars
- Data source: Cargurus, Sept 2020
- 10 gb dataset, sub-sampled to 730k rows
- 66 columns - high dimensionality
- Pre-processing required to extract numeric features

Column Name	Datatype
vin	String
back_legroom	String
bed	String
bed_height	String
bed_length	String
body_type	String
cabin	String
city	String
city_fuel_economy	Float
combine_fuel_economy	Float
daysonmarket	Integer
dealer_zip	String
description	String
engine_cylinders	String
engine_displacement	Float
engine_type	String
exterior_color	String
fleet	Boolean
frame_damaged	Boolean
franchise_dealer	Boolean
franchise_make	String
front_legroom	String
fueltankvolume	String
fuel_type	String
has_accidents	Boolean
height	String
highway_fuel_economy	Float
horsepower	Float
interior_color	String
isCab	Boolean
is_certified	Boolean
is_cpo	Boolean
is_new	Boolean
is_oemcpo	Boolean
latitude	Float
length	String
listed_date	String
listing_color	String
listing_id	Integer
longitude	Float
main_picture_url	String
major_options	String
make_name	String
maximum_seating	String
mileage	Float
model_name	String
owner_count	Float
power	String
price	Float
salvage	Boolean
savings_amount	Integer
seller_rating	Float
sp_id	Float
sp_name	String
theft_title	Boolean
torque	String
transmission	String
transmission display	String
trimId	String
trim_name	String
vehicle_damage_category	Float
wheel_system	String
wheel_system_display	String
wheelbase	String
width	String
year	Integer

Project Goals

- Identify reduced feature subset
- Create a pre-processing pipeline for data cleaning, OHE, and standardization
- Apply ML Models to predict “Price”:
  - Linear Regression
  - XGBoost
  - SVM Regression
- Tune hyper parameters
- Compare performance

Dataset

	vin	back_legroom	bed	bed_height	bed_length	body_type	cabin	city	city_fuel_economy	combine_fuel_economy	daysonmarket	de
0	ZACNJABB5KPJ92081	35.1 in	NaN	NaN	NaN	SUV / Crossover	NaN	Bayamon	NaN	NaN	522	
1	SALCJ2FX1LH858117	38.1 in	NaN	NaN	NaN	SUV / Crossover	NaN	San Juan	NaN	NaN	207	
2	JF1VA2M67G9829723	35.4 in	NaN	NaN	NaN	Sedan	NaN	Guaynabo	17.0	NaN	1233	

dealer_zip	description	engine_cylinders	engine_displacement	engine_type	exterior_color	fleet	frame_damaged	franchise_dealer	franchise_make
00960	[!@@Additional Info@@!]Engine: 2.4L I4 ZERO EV...	I4	1300.0	I4	Solar Yellow	NaN	NaN	True	Jeep
00922	[!@@Additional Info@@!]Keyless Entry,Ebony Mor...	I4	2000.0	I4	Narvik Black	NaN	NaN	True	Land Rover
00969	NaN	H4	2500.0	H4	None	False	False	True	FIAT

make_name	maximum_seating	mileage	model_name	owner_count	power	price	salvage	savings_amount	seller_rating	sp_id	sp_name	theft_title	torque
Jeep	5 seats	7.0	Renegade	NaN	177 hp @ 5,750 RPM	23141.0	NaN	0	2.8	370599.0	Flagship Chrysler	NaN	200 lb-ft @ 1,750 RPM
Land Rover	7 seats	8.0	Discovery Sport	NaN	246 hp @ 5,500 RPM	46500.0	NaN	0	3.0	389227.0	Land Rover San Juan	NaN	269 lb-ft @ 1,400 RPM
Subaru	5 seats	NaN	WRX STI	3.0	305 hp @ 6,000 RPM	46995.0	False	0	NaN	370467.0	FIAT de San Juan	False	290 lb-ft @ 4,000 RPM

EDA Overview

- Univariate and multivariate analysis
  - Categorical, Numeric, Boolean
  - 66 columns
- High dimensionality features
  - color, transmission description, model
  - “Options” = 110k
  - ‘Ext Color’ = 14k

EDA

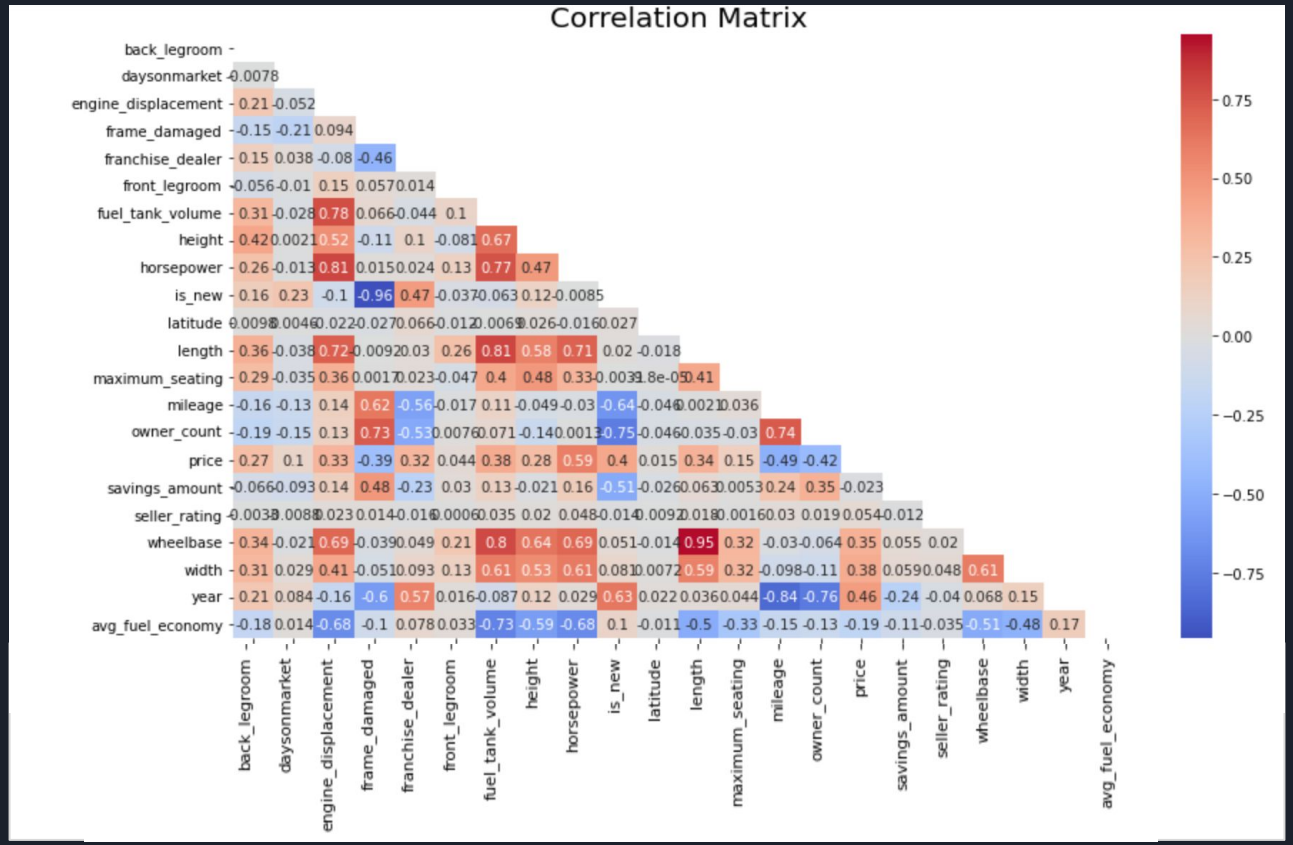
	price	daysonmarket	year	mileage	owner_count	city_fuel_economy	highway_fuel_economy	horsepower
price	1.000000	0.099929	0.455618	-0.485089	-0.415682	-0.168770	-0.243671	0.588620
daysonmarket	0.099929	1.000000	0.084086	-0.130885	-0.145635	0.017604	0.005218	-0.012847
year	0.455618	0.084086	1.000000	-0.842173	-0.763823	0.169868	0.150757	0.028753
mileage	-0.485089	-0.130885	-0.842173	1.000000	0.736786	-0.158930	-0.136023	-0.029965
owner_count	-0.415682	-0.145635	-0.763823	0.736786	1.000000	-0.142355	-0.100897	0.001338
city_fuel_economy	-0.168770	0.017604	0.169868	-0.158930	-0.142355	1.000000	0.934032	-0.660811
highway_fuel_economy	-0.243671	0.005218	0.150757	-0.136023	-0.100897	0.934032	1.000000	-0.684933
horsepower	0.588620	-0.012847	0.028753	-0.029965	0.001338	-0.660811	-0.684933	1.000000

New feature:
average fuel economy

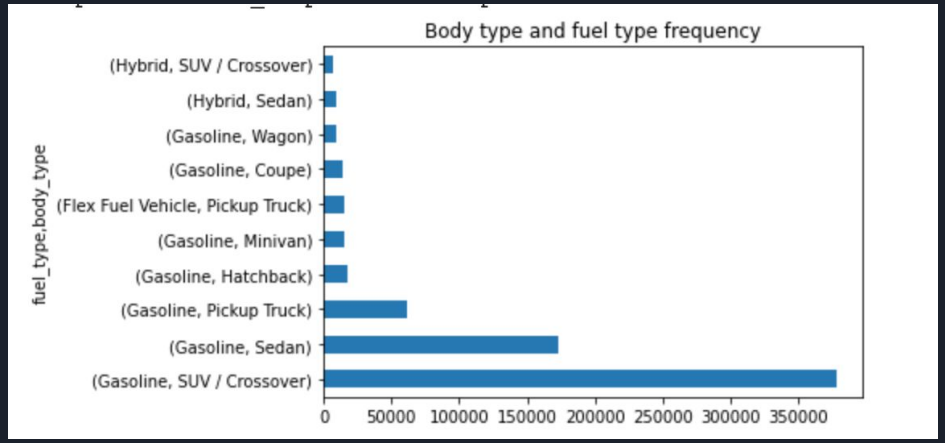
EDA- Correlation Matrix

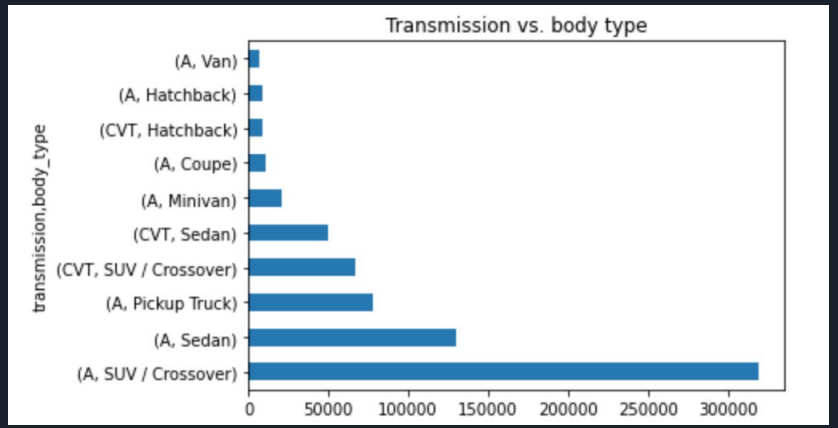
POS
-Horsepower
-Is new
-Fuel tank volume
-Length
-Franchise dealer
-Engine displacement

NEG
-Mileage
-Owner count
-Frame damaged

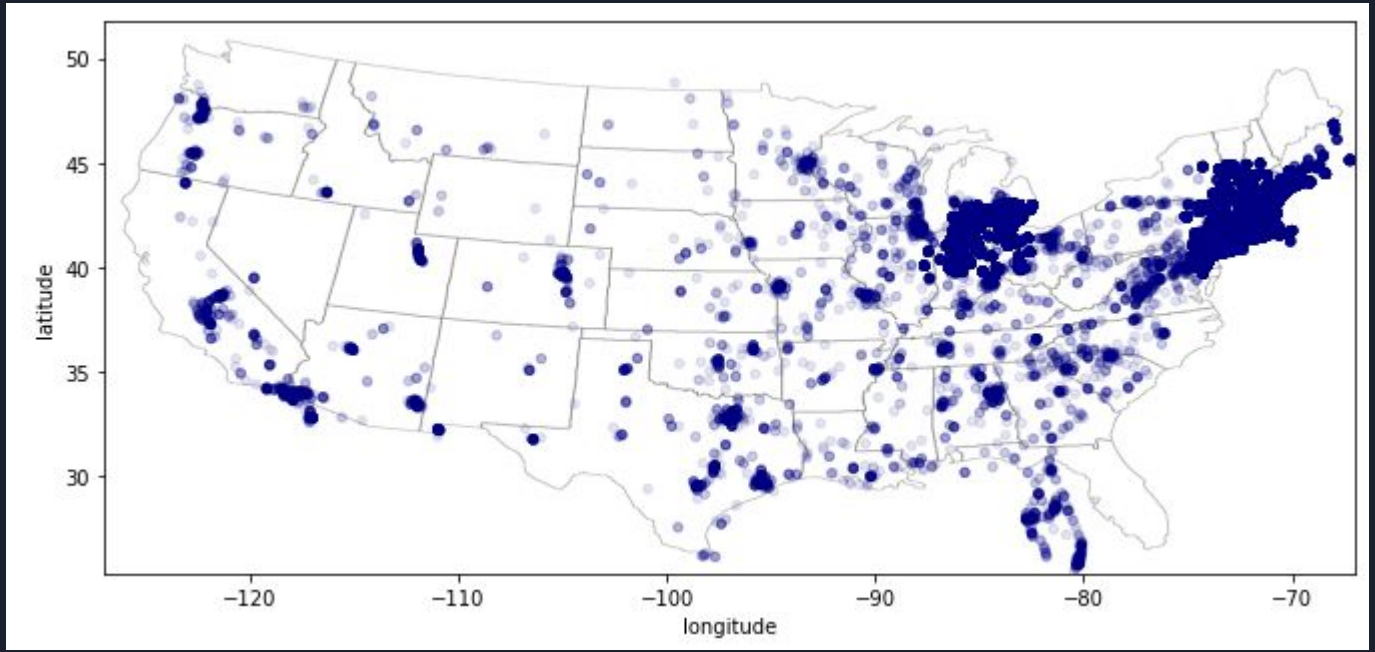


EDA- Multivariate Visualizations





EDA- Visualizations



Data Pre-Processing

- Convert strings → integers
- Null handling
  - Mileage
  - Drop cols with >50% nulls
- Feature engineering/Dimensionality reduction
  - New cols: age, avg fuel economy
- Exported cleaned dataset as parquet
- Test/Train 80/20 split
- OHE and data scaling/standardization using SKlearn pipeline

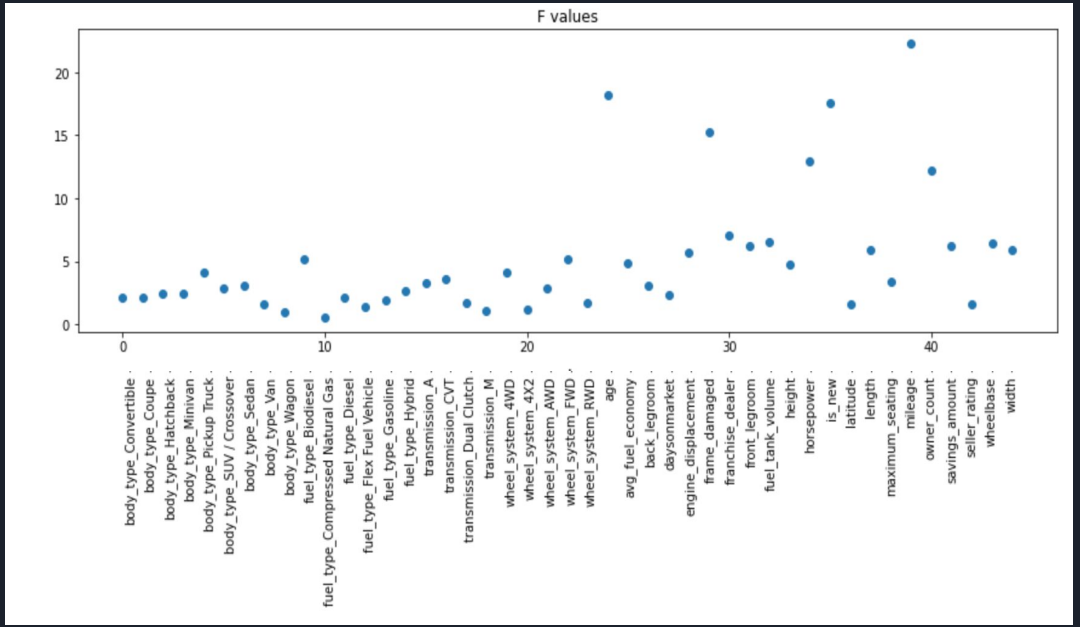
Linear Regression Model

- Baseline: HuberRegressor, alpha = 0.01, epsilon = 1
  - Huber - robust to outliers
  - Train CV MAE = 70.95
  - Test CV MAE = 71.75
- Price ranges from $349 to $2,698,500
  - Train RMSE = 11204
  - Test RMSE = 10441

Feature Importance

High F-values

- Mileage
- Horsepower
- Is new
- Engine displacement
- Age



Feature Selection

- SelectFromModel + LR Model
- threshold = 0.75*mean
  - Identified 10 features:
    - **Age**
    - Avg fuel economy
    - **Engine displacement**
    - Front legroom
    - Fuel tank volume
    - **Horsepower**
    - **Mileage**
    - Owner count
    - Savings amount
    - wheelbase

- Reduced feature model performance scores slightly lower:
  - Train CV MAE = 75.16
  - Test CV MAE = 76.45
- Training time improved

XGBoost Model

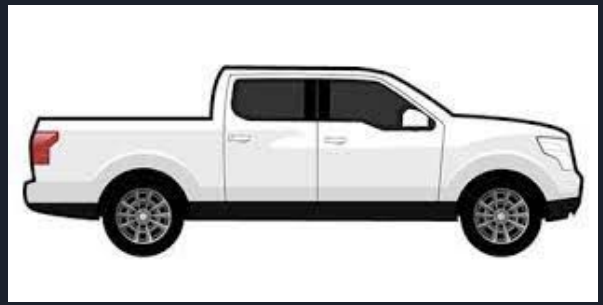
- **Model 1: L2 regularization (reg_lambda = 0.5)**
  - Train Scores
    - RMSE= 4664 , MAE = 2559
    - CV MAE = 51.06
  - Test Scores
    - RMSE = 5414 , MAE = 2608
    - CV MAE = 51.02
- **Model 2: L1 regularization (reg_alpha = 0.5)**
  - Train Scores
    - RMSE = 4681 , MAE = 2544
    - CV MAE =51.17
  - Test Scores
    - RMSE = 5444 , MAE =2589
    - CV MAE = 51.2

```
#1st model for XGBoost
params = {
    'n_jobs':-1,
    'n_estimators':100,
    'max_depth':5,
    'learning_rate':0.3,
    'reg_lambda':0.5   #L2 regularization
}
xgb_reg = XGBRegressor(**params)
xgb_reg.fit(x_train,y_train)
```

```
#2nd model for XGBoost using L1 regularization instead of L2
params = {
    'n_jobs':-1,
    'n_estimators':100,
    'max_depth':5,
    'learning_rate':0.3,
    #'reg_lambda':0.5,  #L2 regularization
    'reg_alpha':0.5     #L1 regularization
                        #for high dimensionality dataset
}
xgb_reg2 = XGBRegressor(**params)
xgb_reg2.fit(x_train,y_train)
```

XGBoost Feature Importance

- Model 1 / Model 2 Highest
  - PIckup Truck body type
  - Horsepower
  - Mileage
  - Flex fuel
  - AWD
  - FWD



SVM Regression Model

- Model 1: Epsilon 0.1
  - Train MAE = 15604.16
  - Test MAE = 15581.77
- RMSE
  - Train RMSE = 26703.70
  - Test RMSE = 25973.58
- Model generalizes well, but overall performance is disappointing.

SVM Regression + Feature Selection

- SelectFromModel +SVM Model
- threshold = 0.75*mean
  - Identified 23 features:
    - 'body_type_Hatchback',
    - 'body_type_Minivan',
    - 'body_type_Pickup Truck',
    - 'body_type_SUV / Crossover',
    - 'body_type_Wagon',
    - 'fuel_type_Diesel',
    - 'fuel_type_Flex Fuel Vehicle',
    - 'fuel_type_Gasoline',
    - 'transmission_CVT',
    - 'transmission_M',
    - 'wheel_system_4WD',
    - 'wheel_system_4X2',
    - 'wheel_system_AWD',
    - 'wheel_system_FWD',
    - 'wheel_system_RWD',
    - 'age',
    - 'frame_damaged',
    - 'franchise_dealer',
    - 'fuel_tank_volume',
    - 'is_new',
    - 'maximum_seating',
    - 'owner_count',
    - 'seller_rating'

- Improved model performance on reduced feature set:
  - Train RMSE = 13684
  - Test RMSE = 13105
  - Train MAE = 6473.59
  - Test MAE = 6503.14
- Training time improved
- Model continues to generalize well and performance vastly improved.
- CV not performed due to long training time and failure to converge with default iterations

Comparison

1. XGBoost Model 1 had the best all around scores, but longest training time
2. Features of importance across models were also the most correlated to price
    a. Horsepower (0.59)
    b. Mileage (-0.49)
    c. Age/Year (0.46)
    d. Pickup Truck (0.20)

Model	Best MAE
Linear Regression	71.75
XGBoost	51.02
SVM Regression w/ feature selection	6503*

Conclusion & Future Works

- Removal of price outliers: focus on “newer typical” used car market improved performance and MAE
- Increase dataset size
- Advanced feature selection, PCA, dimensionality reduction

	price			
	mean	min	max	count
age				
1	20976.354839	495.0	599000.0	682
2	8973.505638	349.0	3195000.0	37962
3	14860.983726	1699.0	1299950.0	83570
4	31880.541226	4990.0	2698500.0	608007

Workload Distribution

- EDA and Pre-Processing: all members
- Linear model- Katy
- XGBoost- Jacob
- SVR - John