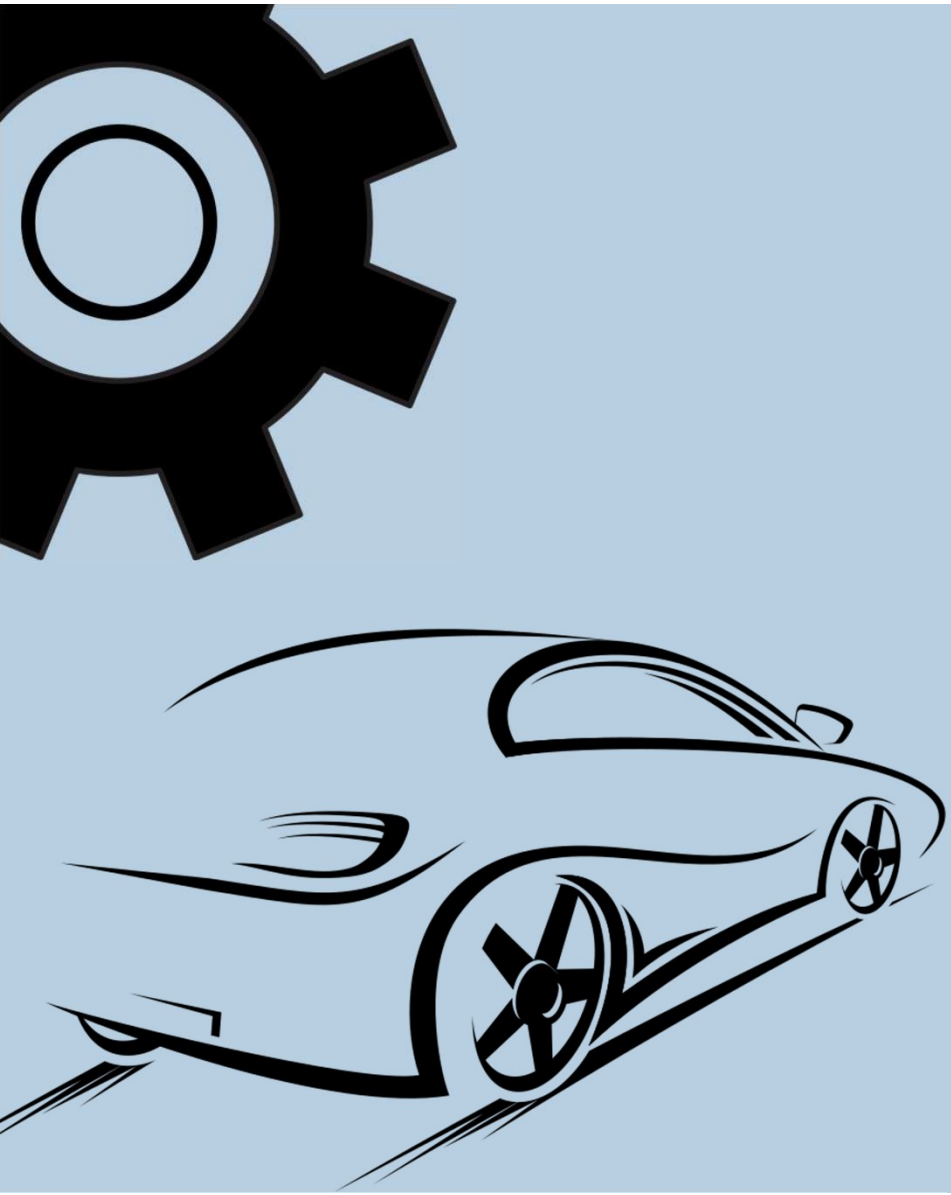


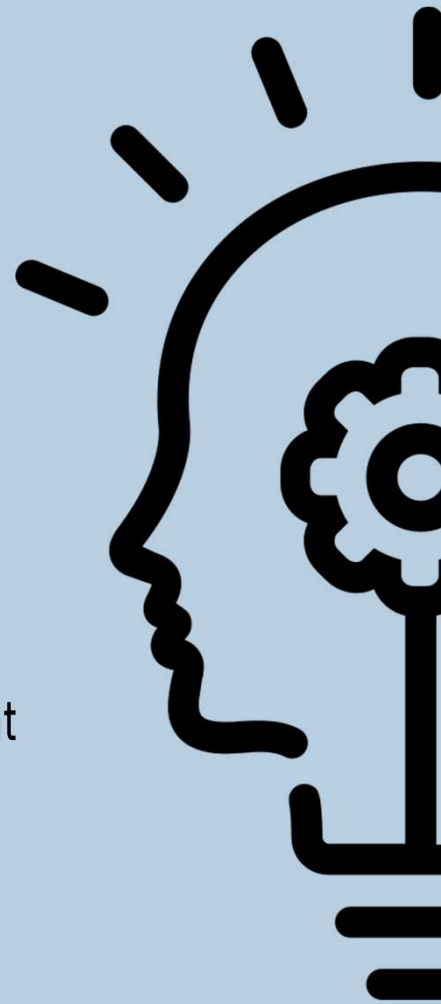
Predictive Analysis for Prices of Used Cars

GROUP 7: ANKIT JAIN, CASEY KAN, DIDO CHANG, SIJIA LI, SONG HAN

Data and Business Ideas

Data

- Listings of Car Sales from craigslist
- Includes variety of attributes associated with the car and including the price of the car.



Business Ideas

Aim to study these 3 main questions through our modelling:

- What is the degree of influence of different attributes towards the class -- price?
- What recommendations can we provide to future buyers when looking for cars that they may possibly buy?
- What recommendations can we provide to future sellers when making appropriate quotes for the cars they sell?

Data Cleaning and Preprocessing

- Used a combination of Weka and Python Programming
- **Initial Data**: 25 columns and 423857 rows
- Removed Columns which were unique to every row
- Removed Rows with missing values
- Removed Rows with outlier values for continuous attributes [price and odometer distance]
- Removed Rows for models dated before 2000
- Removed Columns which created noise
- **Cleaned Data**: 13 columns and 53390 rows
- Discretized Price and Odometer attribute into 3 bins [Frequency and Size of bins varies according to model]

Unsupervised Learning: Association Rules

- Metric Considered: Lift
- Minimum Lift: 1

Trial 1: Price Discretized to 3 bins of equal size with non-equal frequency

{condition_attribute} -> {price_attribute}

Price_Attribute	Condition_Attribute	Lift	Confidence	Number of Matched Rules
Mid Price Range	odometer='(-inf-91000.5]'	1.55	0.43	27
Mid Price Range	drive=4wd	1.27	0.35	27
Mid Price Range	condition=excellent	1.18	0.32	27
Mid Price Range	size=full-size	1.14	0.31	27
Mid Price Range	transmission=automatic	1.01	0.28	81

Price_Attribute	Condition_Attribute	Lift	Confidence	Number of Matched Rules
Lowest Price Range	odometer='(141002.5-inf)'	1.30	0.84	265
Lowest Price Range	drive=fwd	1.21	0.78	633
Lowest Price Range	type=sedan	1.20	0.77	303
Lowest Price Range	condition=good	1.19	0.77	197
Lowest Price Range	size=compact	1.19	0.76	5
Lowest Price Range	paint_color=silver	1.11	0.71	19
Lowest Price Range	cylinders=4_cylinders	1.10	0.71	383
Lowest Price Range	odometer='(91000.5-141002.5]'	1.08	0.69	153
Lowest Price Range	size=mid-size	1.08	0.70	163
Lowest Price Range	cylinders=6_cylinders	1.04	0.67	217
Lowest Price Range	fuel=gas	1.02	0.66	1169
Lowest Price Range	title_status=clean	1.01	0.65	1157

Unsupervised Learning: Association Rules

Trial 2: Price Discritized to 3 bins of non-equal size with equal frequency

{condition_attribute} -> {price_attribute}

Price_Attribute	Condition_Attribute	Lift	Confidence	Number of Matched Rules
Mid Price Range	odometer='(91000.5-141002.5]'	1.30	0.43	27
Mid Price Range	cylinders=4_cylinders	1.15	0.38	41
Mid Price Range	drive=fwd	1.12	0.37	43
Mid Price Range	type=sedan	1.11	0.37	19
Mid Price Range	condition=excellent	1.10	0.36	29
Mid Price Range	type=SUV	1.04	0.34	1
Mid Price Range	size=mid-size	1.04	0.35	19
Mid Price Range	transmission=automatic	1.01	0.33	153
Mid Price Range	fuel=gas	1.01	0.34	153

Price_Attribute	Condition_Attribute	Lift	Confidence	Number of Matched Rules
Lowest Price Range	odometer='(141002.5-inf)'	1.63	0.54	27
Lowest Price Range	condition=good	1.38	0.46	27
Lowest Price Range	drive=fwd	1.28	0.43	75
Lowest Price Range	type=sedan	1.27	0.42	57
Lowest Price Range	size=mid-size	1.11	0.37	27
Lowest Price Range	cylinders=6_cylinders	1.11	0.37	27
Lowest Price Range	cylinders=4_cylinders	1.05	0.35	45
Lowest Price Range	fuel=gas	1.03	0.35	203
Lowest Price Range	title_status=clean	1.02	0.34	195

Price_Attribute	Condition_Attribute	Lift	Confidence	Number of Matched Rules
Max Price Range	odometer='(-inf-91000.5]'	1.73	0.58	49
Max Price Range	cylinders=8_cylinders	1.48	0.49	5
Max Price Range	drive=4wd	1.38	0.46	59
Max Price Range	size=full-size	1.22	0.41	75
Max Price Range	condition=excellent	1.18	0.39	69
Max Price Range	type=SUV	1.09	0.36	9
Max Price Range	transmission=automatic	1.01	0.34	135

Correlation Analysis

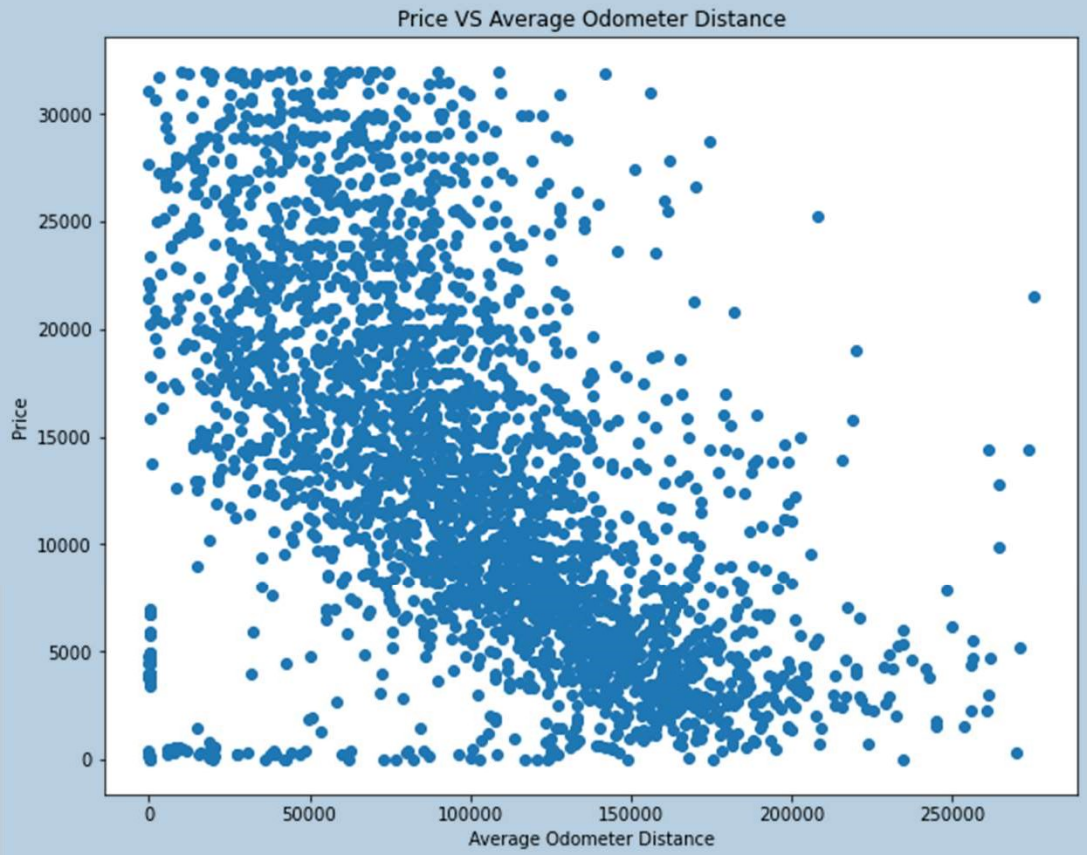
Attribute: Price

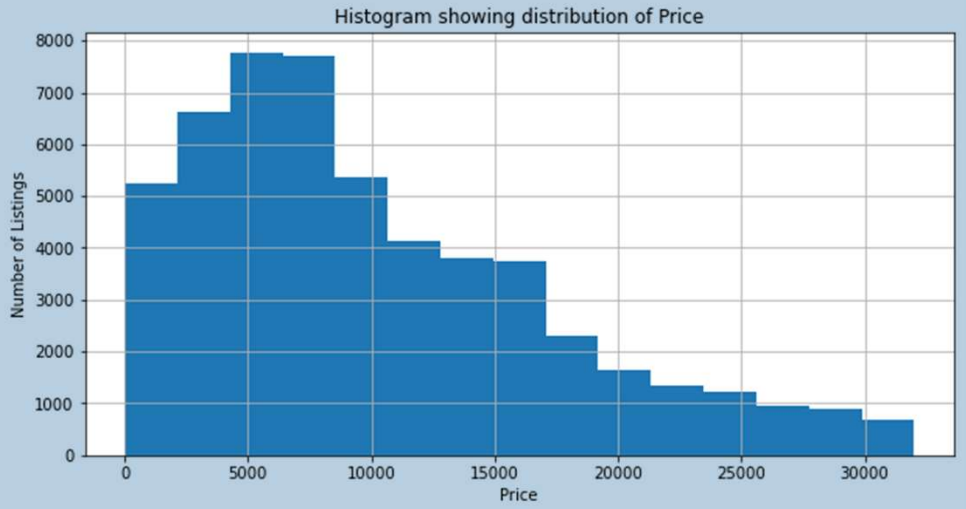
Top 3 Correlated Attributes

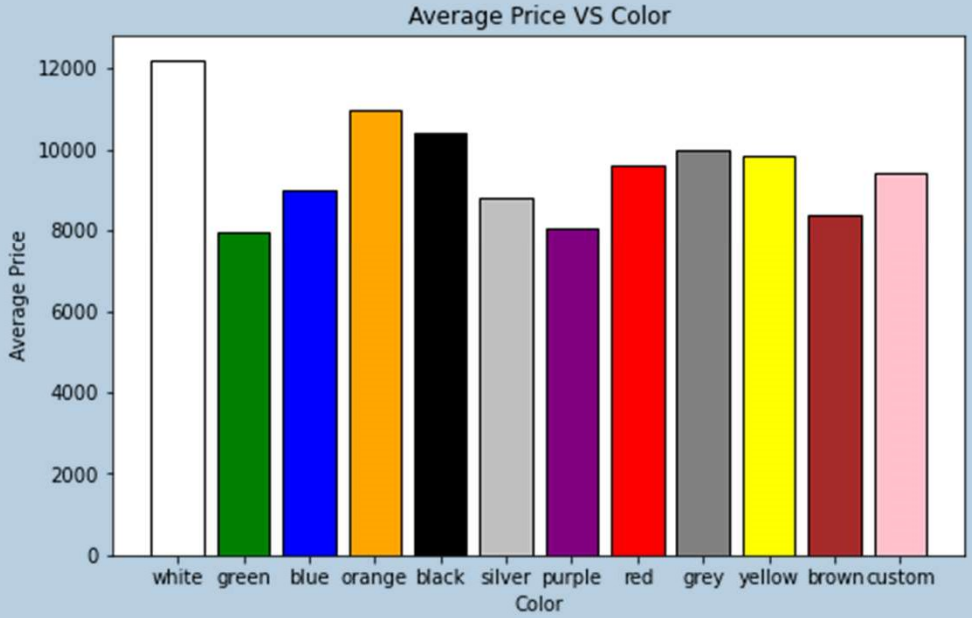
- Odometer [Negative Correlation]: -0.4335
- Drive [Positive Correlation]: 0.2686
- Fuel [Positive Correlation]: 0.1827

```
Ranked attributes:
 0.2686   10 drive
 0.1827    6 fuel
 0.1729   11 size
 0.1499    4 condition
 0.1439   12 type
 0.1424    5 cylinders
 0.1109    2 year
 0.0783   13 paint_color
 0.0625    3 manufacturer
 0.0545    9 transmission
 0.0196    8 title_status
-0.4335    7 odometer
```

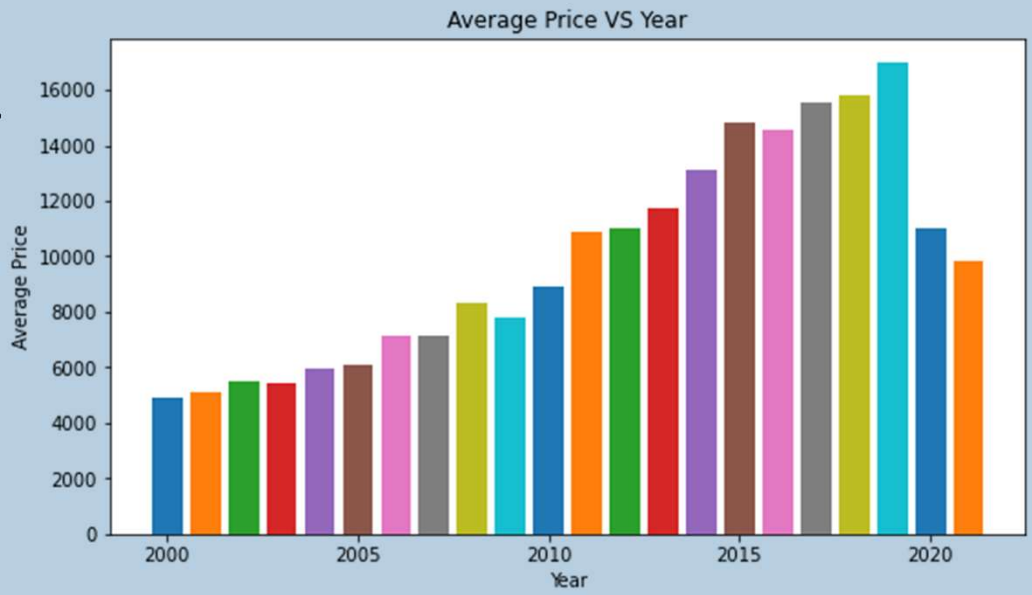
Graphical Analysis

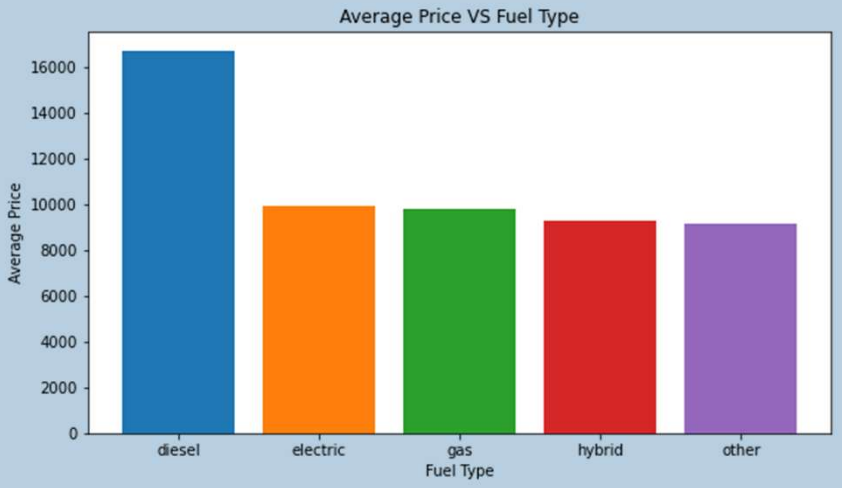


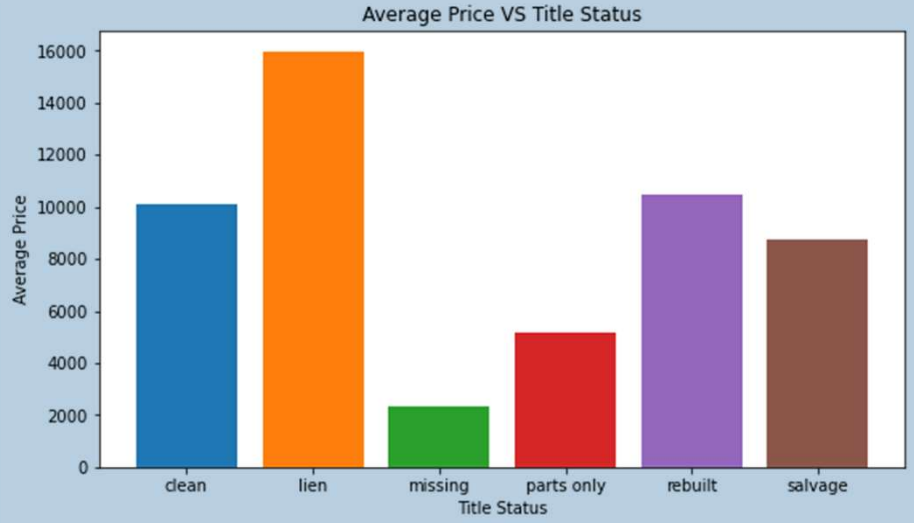




Graphical Analysis







Supervised Learning: Naive Bayes

Dependent Variable: Price

- Price discretized into 3 bins of equal size
- Odometer discretized into 3 bins of equal frequency
- Duplicates Rows Removed
- Accuracy: 76.13% [Percentage Split 66% Test Option]

```
=== Summary ===

Correctly Classified Instances       10873               76.1308 %
Incorrectly Classified Instances      3409               23.8692 %
Kappa statistic                          0.5148
Mean absolute error                      0.2131
Root mean squared error                  0.3273
Relative absolute error                 63.6573 %
Root relative squared error             80.0103 %
Total Number of Instances            14282

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.880    0.267    0.857      0.880    0.868      0.620    0.900     0.946     '(-inf-10666.333333]'
                 0.597    0.155    0.593      0.597    0.595      0.441    0.841     0.594     '(10666.333333-21332.666667]'
                 0.372    0.034    0.488      0.372    0.422      0.383    0.910     0.444     '(21332.666667-inf)'
Weighted Avg.    0.761    0.218    0.755      0.761    0.757      0.552    0.884     0.809

=== Confusion Matrix ===

    a    b    c   <-- classified as
 8109  945  161 |    a = '(-inf-10666.333333]'
 1288 2334  290 |    b = '(10666.333333-21332.666667]'
   66  659  430 |    c = '(21332.666667-inf)'
```

Supervised Learning: Naive Bayes

Contingency table and probability chart

1. Odometer

Count of odometer	Column Labels			
Row Labels	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''	Grand Total
'\'(141002.5-inf)\''	11758	1932	289	13979
'\'(91000.5-141002.5]\''	9718	3620	676	14014
'\'(-inf-91000.5]\''	5619	5967	2427	14013
Grand Total	**27095**	**11519**	**3392**	**42006**

Probability			
Odometer/Price	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''
'\'(141002.5-inf)\''	0.43	0.17	0.09
'\'(91000.5-141002.5]\''	0.36	0.31	0.20
'\'(-inf-91000.5]\''	0.21	0.52	0.72

2. Drive

Count of drive	Column Labels			
Row Labels	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''	Grand Total
4wd	8223	5553	2179	15955
fwd	14869	3789	431	19089
rwd	4003	2177	782	6962
Grand Total	**27095**	**11519**	**3392**	**42006**

Probability			
Drive/Price	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''
4wd	0.30	0.48	0.64
fwd	0.55	0.33	0.13
rwd	0.15	0.19	0.23

3. Fuel

Count of fuel	Column Labels			
Row Labels	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''	Grand Total
diesel	450	746	410	1606
electric	9	4	1	14
gas	26122	10559	2956	39637
hybrid	477	189	22	688
other	37	21	3	61
Grand Total	**27095**	**11519**	**3392**	**42006**

Probability			
Fuel/Price	'\'(-inf-10666.333333]\''	'\'(10666.333333-21332.666667]\''	'\'(21332.666667-inf)\''
diesel	0.02	0.06	0.12
electric	0.00	0.00	0.00
gas	0.96	0.92	0.87
hybrid	0.02	0.02	0.01
other	0.00	0.00	0.00

Supervised Learning: Decision Trees

Dependent Variable: Price

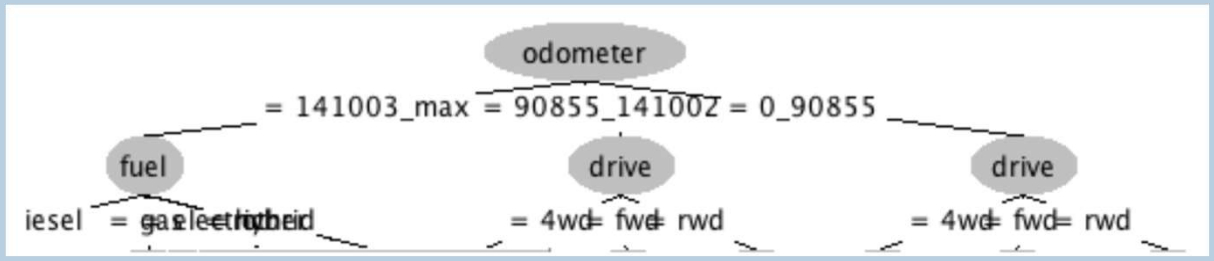
- Price discretized into 3 bins of equal size
- Odometer discretized into 3 bins of equal frequency
- Duplicates Rows NOT removed
- Accuracy: 80.99% [Percentage Split 66% Test Option]
- First Node Split on Odometer Variable
- Improved Accuracy from Naive Bayes Model

```
=== Summary ===

Correctly Classified Instances       14702               80.9894 %
Incorrectly Classified Instances      3451               19.0106 %
Kappa statistic                          0.641
Mean absolute error                      0.1711
Root mean squared error                  0.314
Relative absolute error                 48.2972 %
Root relative squared error             74.4099 %
Total Number of Instances            18153

=== Confusion Matrix ===

    a    b    c   <-- classified as
 9796 1069  177 |    a = '(-inf-10666.333333]'
 1234 3816  283 |    b = '(10666.333333-21332.666667]'
  132  556 1090 |    c = '(21332.666667-inf)'
```



Supervised Learning: Random Forests

Dependent Variable: Price

- Price discretized into 3 bins of equal size
- Odometer discretized into 3 bins of equal frequency
- Duplicates Rows NOT removed
- 3 Features selection gave most accurate model
- Accuracy: 84.95% [Percentage Split 66% Test Option]
- Removed Overfitting from Decision Trees
- Highly correlated attributes weren't used rather year, manufacturer and paint_color helped most, so more attributes included

```
Attribute importance based on average impurity decrease (and number of nodes using that attribute)

      0.76 ( 46042)  year
      0.63 ( 65740)  manufacturer
      0.58 ( 90404)  paint_color
      0.52 ( 46887)  type
      0.5  ( 52738)  odometer
      0.48 ( 76137)  condition
      0.47 ( 51734)  cylinders
      0.45 ( 62963)  size
      0.43 ( 48174)  drive
      0.42 ( 23730)  fuel
      0.37 ( 32408)  title_status
      0.36 ( 23289)  transmission
```

```
--- Summary ---

Correctly Classified Instances       15421               84.9501 %
Incorrectly Classified Instances      2732               15.0499 %
Kappa statistic                          0.7146
Mean absolute error                      0.1525
Root mean squared error                  0.2687
Relative absolute error                 43.0464 %
Root relative squared error             63.6709 %
Total Number of Instances            18153

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.918     0.165     0.896      0.918     0.907     0.758     0.952     0.968     '(-inf-10666.333333]'
               0.759     0.096     0.767      0.759     0.763     0.666     0.927     0.840     '(10666.333333-21332.666667]'
               0.696     0.020     0.788      0.696     0.739     0.715     0.969     0.830     '(21332.666667-inf)'
Weighted Avg.  0.850     0.130     0.848      0.850     0.848     0.727     0.946     0.917
```

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
 10134   804   104 |     a = '(-inf-10666.333333]'
  1055  4049   229 |     b = '(10666.333333-21332.666667]'
   116   424  1238 |     c = '(21332.666667-inf)'
```

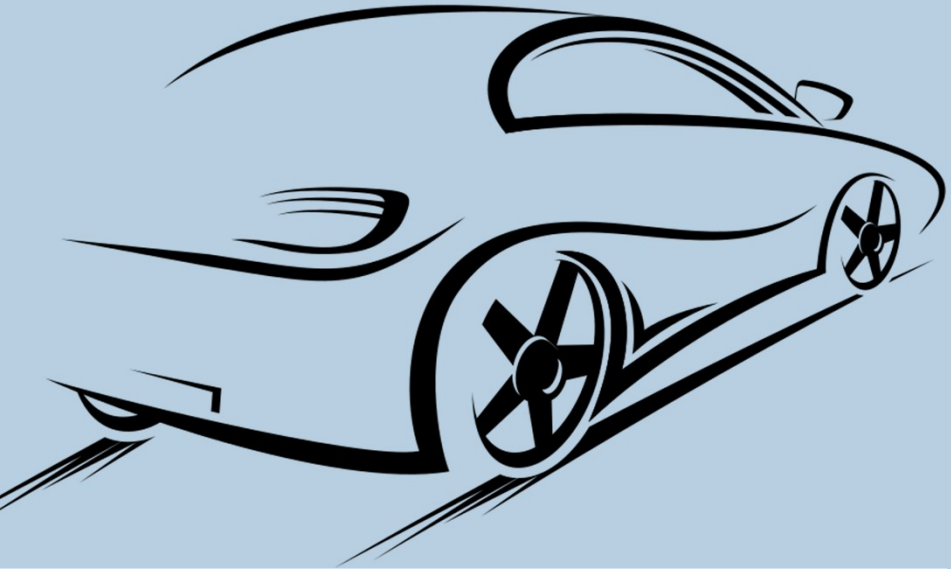
Conclusions and Model Benefits

- Best Performing Model: Random Forests: 84.9501% accuracy
- Major Predictors for Price of Used Car: Odometer Distance, Drive, Fuel Type
- Benefits for seller: set a price range
- Benefits for buyer: set a budget

Limitations and Improvements

- Numerical Data
- Details of Condition of Car and proper regional data
- Limited to only US Market region
- Uneven Distribution of certain attributes
- Seller/buyer information

THANK YOU!



QUESTIONS