

capstone project [part4]

updates

carsales.com.au
Australia's No.1 because it works!

OUTLINE

- PROBLEM STATEMENT
- DATA SCRAPING
- CLEANING & PREPROCESSING
- EXPLORATORY ANALYSIS
- MODELLING
- RESULTS VISUALIZATION

PROBLEM STATEMENT

Car-Buying Decision-Making **Problem:**

	Toyota Corolla	Toyota Corolla	Toyota Corolla	Toyota Corolla
Price (1000 \$)	13.9	6.5	15.99	16.99
Year	2010	2006	2014	2015
Odometer	125,629	190,000	28,855	32,532

DATA SCRAPING


Scraping data from www.carsales.com.au

Car make - Toyota Corolla

4,156 Toyota Corolla Cars For Sale

Sort by: Featured

2010 Toyota Corolla Ascent Sport



```
</div>
  <div class="owl-nav">...</div>
  <div class="owl-dots">...</div>
</div>
<div class="media-icons">...</div>
<ul class="badges">...</ul>
</div>
</div>
...
<div class="n_pad-left-20 n_pad-top-5 n_width-max">
  <div class="vehicle-features">
    <div class="listing-feature n_margin-top-5">
      <div class="feature-title">Odometer</div>
      <div class="feature-text">125,629 km</div>
    </div>
    <div class="listing-feature n_margin-top-5">
      <div class="feature-title">Body</div>
      <div class="feature-text">Hatch</div>
    </div>
    <div class="listing-feature n_margin-top-5">
      <div class="feature-title">Transmission</div>
      <div class="feature-text">Automatic</div>
    </div>
    <div class="listing-feature n_margin-top-5">
      ...</div>
    </div>
  </div>
  <div class="pad-top-5 price-column">...</div>
</div>
<div class="comments">...</div>
<div class="n_columns location-info">...</div>
```



DATA CLEANING & PREPROCESSING

Original data was quite messy

	fuel_efficiency	location	make	price	vechile features
0	11L/100km or less	NaN	2016 Toyota Corolla Ascent Sport Auto	\$24,440</td> <td>Body\r\nHatch\r\n\r\n\r\n\r\nTransmission\r\nAutom...</td> </tr> <tr> <th>1</th> <td>11L/100km or less</td> <td>NaN</td> <td>2016 Toyota Corolla Ascent Sport Auto</td> <td>\$24,440	
					Body\r\nHatch\r\n\r\n\r\n\r\nTransmiss

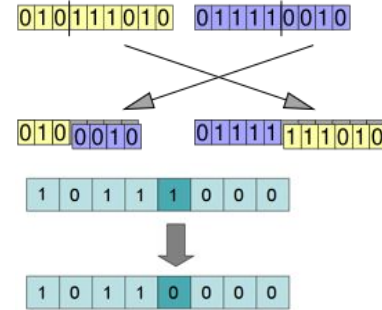
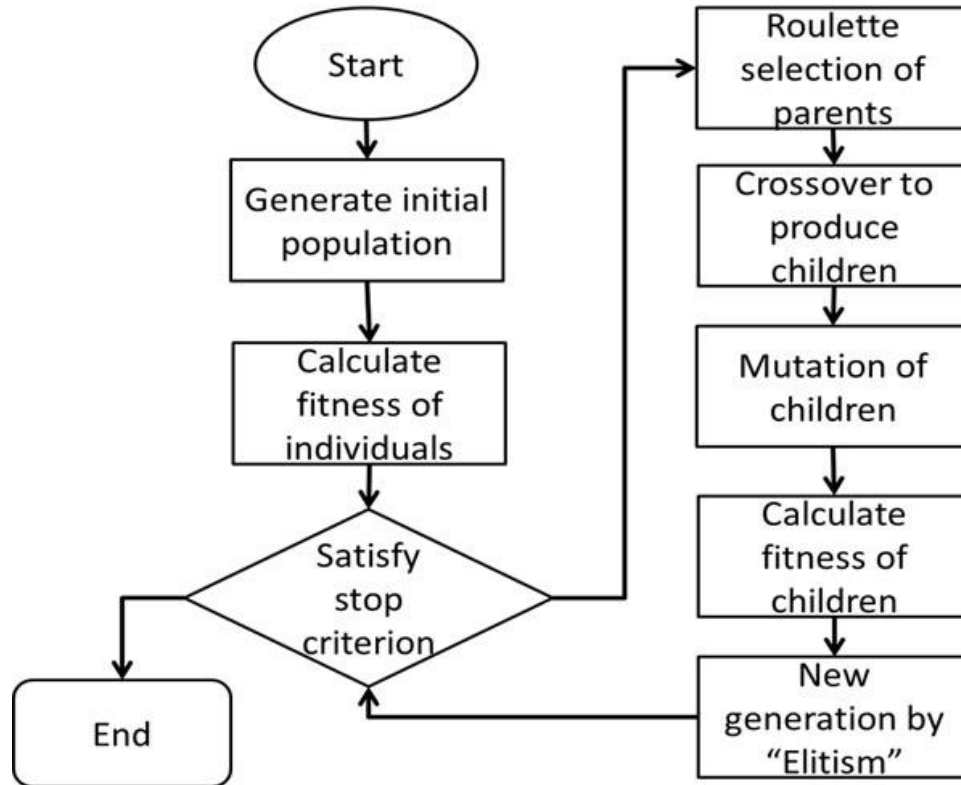
After cleaning Final Data set was 3397

GENETIC ALGORITHMS

In [computer science](#) and [operations research](#), a **genetic algorithm (GA)** is a [metaheuristic](#) inspired by the process of [natural selection](#) that belongs to the larger class of [evolutionary algorithms](#) (EA). Genetic algorithms are commonly used to **generate high-quality solutions to optimization and search problems** by relying on bio-inspired operators such as [mutation](#), [crossover](#) and [selection](#). [WIKI]

- A genetic algorithm is a search heuristic that mimics the process of natural evolution.
- There are five phases
 - Initial Population
 - Fitness Function
 - Selection
 - Crossover
 - Mutation
- The primary advantage of GA's comes from the crossover operation.

GENETIC ALGORITHMS



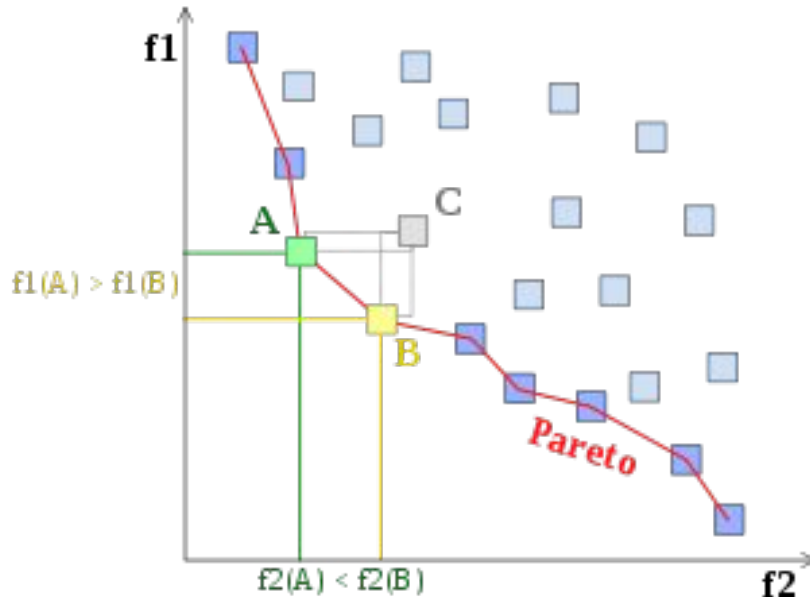
the best organism(s) from the current generation to carry over to the next, unaltered

NSGA-II. PyBRAIN library

Non-dominated Sorting Genetic Algorithm-II (NSGA-II)



Solutions A and B are non-dominated solutions.



Definition 3.1 A solution $\mathbf{x}^{(1)}$ is said to dominate the other solution $\mathbf{x}^{(2)}$, if both the following conditions are true:

1. The solution $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives. Thus, the solutions are compared based on their objective function values (or location of the corresponding points ($\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$) on the objective space).
2. The solution $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective.

For a given set of solutions (or corresponding points on the objective space, for example, those shown in Figure 5(a)), a pair-wise comparison can be made using the above definition and whether one point dominates the other can be established. All points which are not dominated by any other member of the

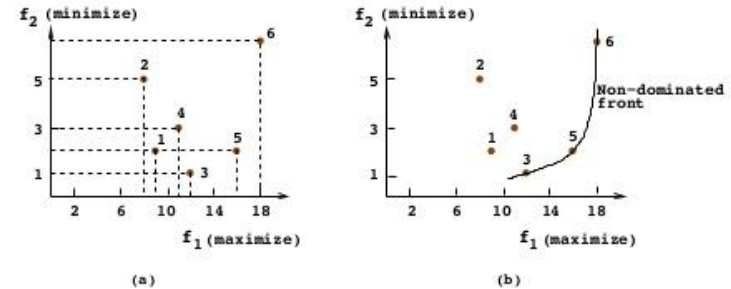
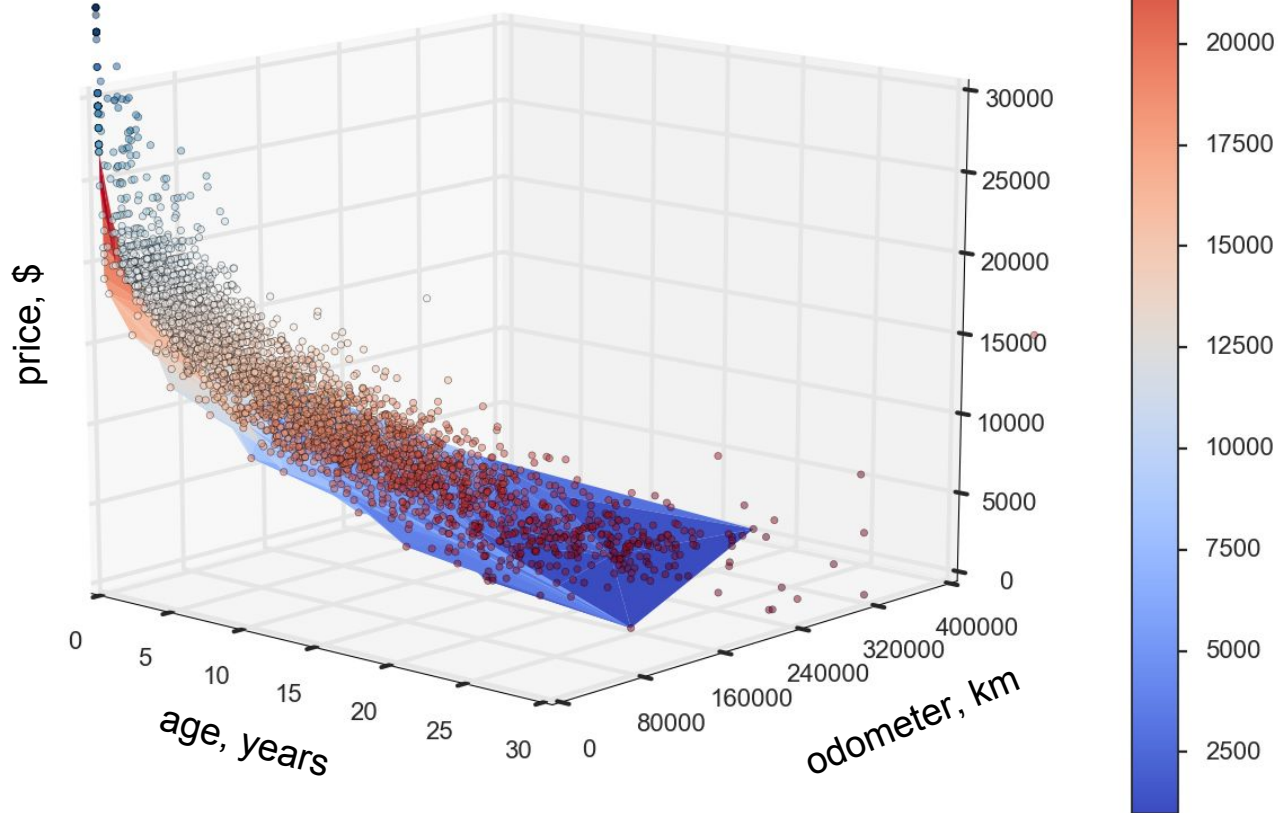


Figure 5: A set of points and the first non-domination front are shown.

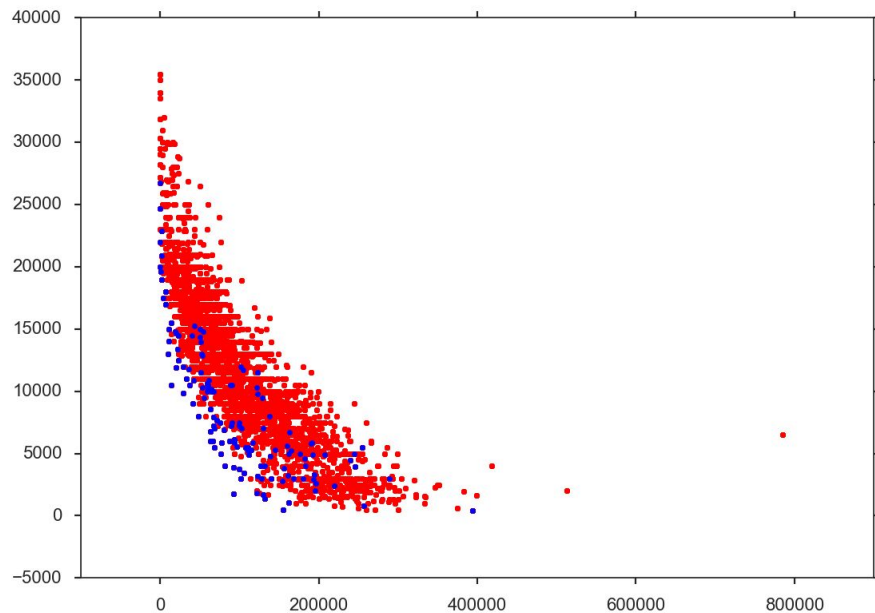
3D PLOT



133 non-dominated
solutions (out of 3397)

Epsilon resolution $1e-9$

2D PLOT



LIBRARIES USED

- DATA SCRAPING: BeautifulSoup, urllib2
- CLEANING & PREPROCESSING: Regular Expressions, Numpy, Pandas
- EXPLORATORY ANALYSIS: Matplotlib, SKlearn
- MODELLING: NSGA-II
- RESULTS VISUALIZATION: Basemap, Matplotlib, Plotly

REFERENCES

1. Kalyanmoy Deb (23 March 2009). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons. ISBN 978-0-470-74361-4. Retrieved 1 November 2012.
- 2.

DATA changes

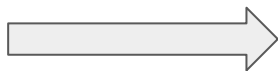
SCRAPING AGAIN....

.....

.....

CLEANING AGAIN

.....



clean data set with 3397 cars

	fuel_efficiency	price	year	odometer	Number of cylinders	capacity
count	3397.0	3397.000000	3397.000000	3397.000000	3397.0	3397.000000
mean	11.0	12712.374154	2009.443921	93639.964086	4.0	1.793318
std	0.0	6325.436554	5.203926	69126.368582	0.0	0.037195
min	11.0	400.000000	1984.000000	1.000000	4.0	1.300000
25%	11.0	7990.000000	2007.000000	41500.000000	4.0	1.800000
50%	11.0	12888.000000	2011.000000	76801.000000	4.0	1.800000
75%	11.0	16990.000000	2014.000000	133600.000000	4.0	1.800000
max	11.0	35449.000000	2016.000000	785500.000000	4.0	2.000000