Auto-MPG Summary

Methodology

- Each model was run with each type of feature selection used in both languages.
- In the symbolic models and quadratic models in Scala, powers of -2, -1 and 1 were used as a base. Several other powers were initially tested but the aforementioned powers were found to generally provide the highest R^2 values.
- For the Symbolic Regression and Symbolic Ridge regression in Scala, all the combinations of having an intercept or the first cross term present were tested. The third cross term was always set to *false* for length of processing the model as well as a way to reduce the number of combinations to test for each type of feature regression
- Symbolic Lasso Regression in Scala was not tested for this data set as the length of code processing time could happen to exceed 20 minutes per test of each combination of parameters with each selection feature method.
- Neither Symbolic Lasso or Symbolic Ridge regression packages could be found in Python and thus were not used.
- Stepwise Selection package could not be found in Python instead Recursive Feature Selection was tested.
- The original file was edited into “auto-mpg-clean.csv” to be read into Scala for easier access, i.e. the response variable was moved to the last column and missing data rows already removed.

Feature Indexes

- 0- Cylin
- 1 - Displacement
- 2 - Horsepower
- 3 - Weight
- 4 - Acceleration
- 5 - Model Year
- 6 - Origin
- 7 - mpg

Table 1. auto-mpg.csv Forward Selection Model Results in Scala.

Forward Selection Models	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.815088	0.812206	0.83578	(0, 5, 3, 6, 2, 1, 4)
Ridge	0.815088	0.811717	0.83578	(0, 5, 3, 6, 2, 1, 4)
Lasso	0.815088	0.812206	0.83578	(0, 5, 3, 6, 2, 1, 4)
Quadratic	0.868855	0.863985	0.853663	(0, 5, 3, 10, 11, 4, 12, 2, 9, 6, 13, 1, 8, 7
Symbolic(true, false)	0.894155	0.888147	0.873931	0, 11, 6, 20, 10, 21, 12, 13, 9, 16, 2, 15, 8, 1, 3, 14, 18, 4, 19, 5, 17, 7
Symbolic(false, true)	0.907593	0.896768	0.873536	(0, 10, 31, 9, 18, 40, 5, 12, 19, 14, 7, 8, 29, 28, 15, 2, 33, 27, 11, 6, 35, 41, 4, 38, 13, 20, 16, 22, 21, 32, 26, 34, 3, 24, 30, 39, 37, 23, 1, 17, 25, 36)
Symbolic (true, true)	0.907557	0.896432	0.87394	0, 11, 6, 20, 10, 21, 12, 13, 32, 1, 15, 8, 9, 30, 29, 18, 25, 41, 14, 19, 36, 28, 5, 42, 39, 35, 17, 22, 24, 40, 33, 16, 31, 2, 34, 38, 23, 4, 3, 26, 27, 37, 7
Symbolic (false, false)	0.894206	0.888503	0.873137	(0, 10, 5, 12, 19, 16, 20, 11, 9, 2, 14, 7, 8, 6, 13, 1, 17, 3, 18, 4, 15)
Symbolic Ridge (true, false)	0.890195	0.87698	0.863801	(0, 5, 3, 24, 40, 29, 6, 26, 33, 7, 2, 14, 39, 20, 34, 22, 41, 23, 25, 27, 4, 1, 32, 31, 35, 37, 28, 30, 36, 13, 16, 11, 19, 15, 9, 12, 17, 8, 10, 18, 21, 38)
Symbolic Ridge (false, true)	0.900507	0.883875	0.852874	(0, 5, 3, 22, 1, 7, 39, 4, 35, 14, 55, 26, 6, 18, 29, 2, 36, 52, 32, 40, 21, 27, 48, 54, 50, 38, 23, 25, 45, 33, 37, 24, 28, 31, 34, 47, 41, 42, 49, 51, 43, 46, 30, 44, 53, 20, 15, 11, 19, 16, 13, 8, 17, 12, 9, 10)
Symbolic Ridge (true, true)	0.917528	0.897304	0.85179	0, 5, 3, 24, 76, 70, 63, 55, 71, 7, 18, 41, 38, 72, 65, 35, 58, 31, 39, 27, 14, 69, 46, 21, 37, 52, 75, 74, 40, 61, 4, 34, 36, 6, 62, 1, 22, 53, 48, 30, 59, 28, 32, 26, 2, 29, 47, 60, 33, 73, 68, 25, 44, 45, 51, 67, 54, 42,

				56, 23, 57, 49, 50, 43, 66, 64, 11, 13, 15, 20, 19, 16, 8, 12, 9, 17, 10
Symbolic Ridge (false, false)	0.831904	0.822363	0.813273	(0, 5, 3, 20, 6, 16, 13, 1, 15, 2, 14, 18, 4, 7, 11, 9, 17, 19, 8, 10, 12

*For the Symbolic the booleans refer to (intercept, cross) and for Symbolic Ridge (cross, cross3).

Table 2. auto-mpg.csv Backward Selection Model Results in Scala.

Backward Selection Models	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.815088	0.812206	0.83578	(0,5)
Ridge	0.815088	0.811717	0.83578	(0,5)
Lasso	0.815088	0.812206	0.83578	(0,5)
Quadratic	0.868855	0.863985	0.853663	(0,5)
Symbolic(true, false)	0.894158	0.888151	0..872697	(0,11)
Symbolic(false, true)	0.907553	0.896428	0.87317	(0,35)
Symbolic (true, true)	0.907593	0.896768	0.873536	(0,6)
Symbolic (false, false)	0.894206	0.888503	0.873137	(0,10)
Symbolic Ridge (true, false)	0.890195	0.87698	0.863801	(0,5)
Symbolic Ridge (false, true)	0.90507	0.883875	0.852874	(0,5)
Symbolic Ridge (true, true)	0.917528	0.897304	0.85179	(0,5)
Symbolic Ridge (false, false)	0.831904	0.822363	0.813273	(0,5)

*For the Symbolic the booleans refer to (intercept, cross) and for Symbolic Ridge (cross, cross3).

Table 3. auto-mpg.csv Stepwise Selection Model Results in Scala.

Stepwise Selection Models	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.814715	0.812315	0.805381	(0, 5, 3, 6, 2, 1)
Ridge	0.814715	0.811828	0.805381	(0, 5, 3, 6, 2, 1)
Lasso	0.814715	0.812315	0.805381	(0, 5, 3, 6, 2, 1)
Quadratic	0.868622	0.864819	0.856799	(0, 5, 3, 10, 11, 4, 12, 2, 9, 6, 13)
Symbolic(true, false)	0.890216	0.885837	0.877332	(0, 11, 6, 20, 10, 12, 13, 9, 2, 15, 8, 1, 3, 14, 18, 4)
Symbolic(false, true)	0.897443	0.891915	0.880934	0, 18, 40, 5, 12, 19, 14, 7, 8, 29, 2, 27, 11, 6, 35, 41, 4, 13, 20, 24, 3
Symbolic (true, true)	0.898834	0.89338	0.880845	0, 6, 20, 21, 12, 13, 15, 8, 9, 30,

				29, 25, 41, 14, 19, 36, 28, 5, 42, 39, 35)
Symbolic (false, false)	0.890255	0.885877	0.872931	(0, 10, 5, 12, 19, 16, 20, 11, 9, 14, 7, 8, 6, 13, 1, 17, 3, 2)
Symbolic Ridge (true, false)	0.88438	0.880403	0.871615	(0, 5, 3, 24, 40, 29, 6, 33, 7, 2, 14, 39, 20, 34, 22)
Symbolic Ridge (false, true)	0.881353	0.877272	0.866769	0, 5, 3, 22, 7, 39, 35, 14, 55, 26, 6, 18, 47
Symbolic Ridge (true, true)	0.897624	0.890929	0.871589	(0, 5, 24, 70, 71, 41, 38, 35, 58, 31, 27, 14, 69, 73, 33, 29, 30, 46, 62, 68, 60, 2, 32, 39, 54, 48, 4, 50)
Symbolic Ridge (false, false)	0.897443	0.891915	0.880934	(0, 5, 3, 20, 6, 16, 13, 1, 15, 2, 14)

*For the Symbolic the booleans refer to (intercept, cross) and for Symbolic Ridge (cross, cross3).

Table 4. auto-mpg.csv Forward Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.8047794757	0.5993815192
Ridge	0.8047790164	0.5997507806
Lasso	0.7936259998	0.682951888
Quadratic	0.8466751275	0.7217200793
Symbolic	0.8106620237	0.8636518201

Table 5. auto-mpg.csv Backward Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.8047794757	0.5993815192
Ridge	0.8047790164	0.5997507806
Lasso	0.7936259998	0.682951888
Quadratic	0.8466751275	0.7217200793
Symbolic	0.8106620237	0.8636518201

Table 6. auto-mpg.csv Recursive Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.7266405655	0.3757740068
Ridge	0.726554551	0.3822181579

Lasso	0.7261356194	0.4008274442
Quadratic	0.7306392076	0.5848002019
Symbolic	0.6388156952	0.635885214