Summary SeoulBikeData

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 (positive powers were necessitated as the data set contained non-positive values). 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 input variable of "Functional Day" was eliminated due to every instance being 'Fun' and "Date" was eliminated.
- For "Seasons" variables, each season was given a value (1,2,3,4) for Winter, Spring, Summer, Autumn. For "Holiday", the no or yes response was changed to 1 and 2 respectively. As well "Hour", "Solar Radiation", "Rainfall" and "Snowfall" all had a numerical one added to each instance in order to eliminate zeros for proper use of some of the models.
- The original data file was edited into "Bikes.csv" to be read into Scala for easier access to the dataframe. Changes include: removing column headers, moving response variable to the last column, and the changes mentioned above and in the Report.

Feature Indexes:

- 0 - Hour: Hour of the Day
- 1 - Temperature: Celsius
- 2 - Humidity: Percent
- 3 - Wind Speed: m/s
- 4 - Visibility: 10 m
- 5 - Dew point Temperature: Celsius
- 6 - Solar Radiation: MJ/m^2
- 7 - Rainfall: mm
- 8 - Snowfall: cm
- 9 - Seasons: Winter, Spring, Summer Autumn
- 10 - Holiday: Holiday or No Holiday

Table 1. SeoulBikeData.csv Forward Selection Model Results in Scala.

Model	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.479406	0.478751	0.476579	((0, 1, 7, 5, 6, 9, 2, 3, 8, 4)
Ridge	0.479406	0.478751	0.476579	(0, 1, 7, 5, 6, 9, 2, 3, 8, 4)
Lasso	0.479406	0.478751	0.476579	((0, 1, 7, 5, 6, 9, 2, 3, 8, 4)
Quadratic	0.519639	0.518540	0.513681	(0, 2, 1, 13, 3, 16, 17, 8, 7, 10, 20, 12, 18, 6, 4, 14, 9, 19, 5, 11, 15)
Symbolic(true, false)	0.487282	0.486578	0.485125	(0, 13, 9, 12, 15, 19, 3, 23, 14, 8, 6, 27, 24, 10, 22, 16, 1, 26, 2, 25, 4, 5, 21, 20, 11, 18, 17, 7
Symbolic(false, true)	0.550726	0.549388	0.547682	(0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10, 17, 16, 22, 6)
Symbolic (true, true)	0.551001	0.549612	0.547645	(0, 13, 9, 12, 15, 19, 3, 23, 14, 8, 6, 27, 24, 10, 22, 16, 1, 26, 2, 25, 4, 5, 21, 20, 11, 18, 17, 7
Symbolic (false, false)	0.519640	0.518485	0.513711	(0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5, 14, 4, 8, 10, 18)
Symbolic Ridge (true, false)	0.550726	0.549337	0.547682	(0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10, 17, 16, 22, 6)
Symbolic Ridge (false, true)	0.552541	0.5509	0.548857	(0, 16, 1, 30, 22, 18, 7, 10, 4, 8, 2, 24, 28, 5, 12, 23, 29, 31, 25, 9, 13, 21, 17, 20, 14, 3, 27, 11, 6, 19, 15, 26)
Symbolic Ridge (true, true)	0.568893	0.566567	0.56428	(0, 31, 1, 24, 19, 33, 44, 13, 22, 38, 43, 14, 9, 15, 46, 28, 3, 39, 7, 10, 36, 40, 17, 20, 32, 4, 18, 5, 29, 23, 11, 45, 12, 34, 21, 16, 30, 25, 6, 26, 8, 2, 42, 27, 37, 41, 35)
Symbolic Ridge (false, false)	0.515577	0.514468	0.509301	((0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5, 14, 4, 8, 10, 18)
Symbolic Lasso (true, false)	0.550726	0.549388	0.547682	(0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10, 17, 16, 22, 6)
Symbolic Lasso (false,	0.552541	0.550951	0.548857	(0, 16, 1, 8, 2, 24, 28, 5, 12,

true)				23, 29, 31, 25, 9, 13, 21, 17, 20, 14, 3, 27, 11, 6, 19, 15, 26)
Symbolic Lasso (true, true)	0.568893	0.566617	0.564278	(0, 31, 1, 24, 19, 33, 44, 13, 22, 38, 43, 14, 9, 15, 46, 28, 3, 39, 7, 10, 36, 40, 17, 20, 32, 4, 18, 5, 29, 23, 11, 45, 12, 34, 21, 16, 30, 25, 6, 26, 8, 2, 42, 27, 37, 41, 35)
Symbolic Lasso (false, false)	0.519369	0.518485/	0.513681	0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5)

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

Table 2. SeoulBikeData.csv Backward Selection Model Results in Scala.

Model	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.479406	0.478751	0.4751326	(0, 1)
Ridge	0.479406	0.478751	0.476579	0, 1)
Lasso	0.479406	0.478751	0.4751326	(0, 1)
Quadratic	0.519639	0.518540	0.513681	(0,12)
Symbolic(true, false)	0.487282	0.486578	0.485125	(0,2)
Symbolic(false, true)	0.550726	0.549388	0.547682	(0,11)
Symbolic (true, true)	0.551001	0.549612	0.547645	(0, 12)
Symbolic (false, false)	0..519937	0.518783	0.510607	(0,1)
Symbolic Ridge (true, false)	0.550726	0.549337	0.547682	(0,11)
Symbolic Ridge (false, true)	0.552541	0.5509	0.548857	(0,11)
Symbolic Ridge (true, true)	0.568893	0.566567	0.56428	(0,11)
Symbolic Ridge (false, false)	0.519639	0.518430	0.513682	(0,1)
Symbolic Lasso (true, false)	0.550726	0.549388	0.547682	(0,11)
Symbolic Lasso (false, true)	0.552541	0.550951	0.548857	(0,11)
Symbolic Lasso (true, true)	0.568893	0.566617	0.564278	(0,11)
Symbolic Lasso (false, false)	0.596397	0.5148485	0.5013681	(0,11)

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

Table 3. SeoulBikeData.csv Stepwise Selection Model Results in Scala.

Model	R2	Adjusted R2	Cross Validated R2	Features Used
Linear Regression	0.479406	0.478751	0.476579	((0, 1, 7, 5, 6, 9, 2, 3, 8)

Ridge	0.479406	0.478751	0.476579	(0, 1, 7, 5, 6, 9, 2, 3, 8)
Lasso	0.479406	0.478751	0.476579	((0, 1, 7, 5, 6, 9, 2, 3, 8)
Quadratic	0.519638	0.518649	0.514103	(0, 2, 1, 13, 3, 16, 17, 8, 7, 10, 20, 12, 18, 6, 4, 14, 9, 19, 5)
Symbolic(true, false)	0.48727	0.486684	0.485449	(0, 2, 1, 9, 3, 12, 5, 6, 7, 10, 4)
Symbolic(false, true)	0.550596	0.549464	0.547925	0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10
Symbolic (true, true)	0.550283	0.549254	0.547598	(0, 13, 9, 12, 15, 19, 3, 14, 8, 6, 27, 24, 10, 22, 16, 1, 26, 2, 25, 4, 5)
Symbolic (false, false)	0.519638	0.518649	0.51403	(0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5
Symbolic Ridge (true, false)	0.550596	0.549412	0.547925	(0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10)
Symbolic Ridge (false, true)	0.551748	0.55062	0.549004	(0, 16, 1, 30, 22, 18, 4, 8, 2, 28, 5, 12, 23, 29, 31, 9, 13, 21, 17, 20, 14, 3)
Symbolic Ridge (true, true)	0.56382	0.564592	0.562821	(0, 31, 1, 24, 19, 13, 38, 43, 14, 9, 15, 46, 28, 3, 39, 7, 10, 36, 40, 20, 4, 18, 5, 29, 23, 11, 45, 12, 34, 21, 16, 30, 25, 6, 37, 27)
Symbolic Ridge (false, false)	0.519638	0.518594	0.514105	((0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5
Symbolic Lasso (true, false)	0.550596	0.549464	0.547925	0, 12, 4, 13, 2, 11, 8, 24, 14, 18, 26, 9, 15, 21, 25, 20, 5, 23, 1, 7, 19, 3, 10
Symbolic Lasso (false, true)	0.551748	0.550671	0.549004	(0, 16, 1, 30, 22, 18, 4, 8, 2, 28, 5, 12, 23, 29, 31, 9, 13, 21, 17, 20, 14, 3)
Symbolic Lasso (true, true)	0.566382	0.564642	0.562821	(0, 31, 1, 24, 19, 13, 38, 43, 14, 9, 15, 46, 28, 3, 39, 7, 10, 36, 40, 20, 4, 18, 5, 29, 23, 11, 45, 12, 34, 21, 16, 30, 25, 6, 37, 27)
Symbolic Lasso (false, false)	0.519638	0.518649	0.514103	(0, 1, 12, 2, 15, 16, 9, 7, 19, 11, 17, 6, 3, 13, 5)

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

Table 4. SeoulBikeData.csv Forward Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.473895595	0.4708007831
Ridge	0.473895594	0.4708012556
Lasso	0.4738906099	0.4708215313
Quadratic	0.5595061581	0.5567926659
Symbolic	0.4121475925	0.4207922428

Table 5. SeoulBikeData.csv Backward Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.473895595	0.4708007831
Ridge	0.473895594	0.4708012556
Lasso	0.4738906099	0.4708215313
Quadratic	0.5595061581	0.5567926659
Symbolic	0.4121475925	0.4207922428

Table 6. SeoulBikeData.csv Recursive Selection Model Results in Python.

Model	Adjusted R2	Cross Validated R2
Linear Regression	0.1905205303	0.1904524628
Ridge	0.1905144878	0.1903996672
Lasso	0.1905133959	0.1903929074
Quadratic	0.2982139355	0.2889846321
Symbolic	0.2487142974	0.2462618413