Data Science Homework 3

R11922066 資工碩一 林庭安

Problem 1 One-by-one Feature Selection

- Feature selection method: Fisher Score

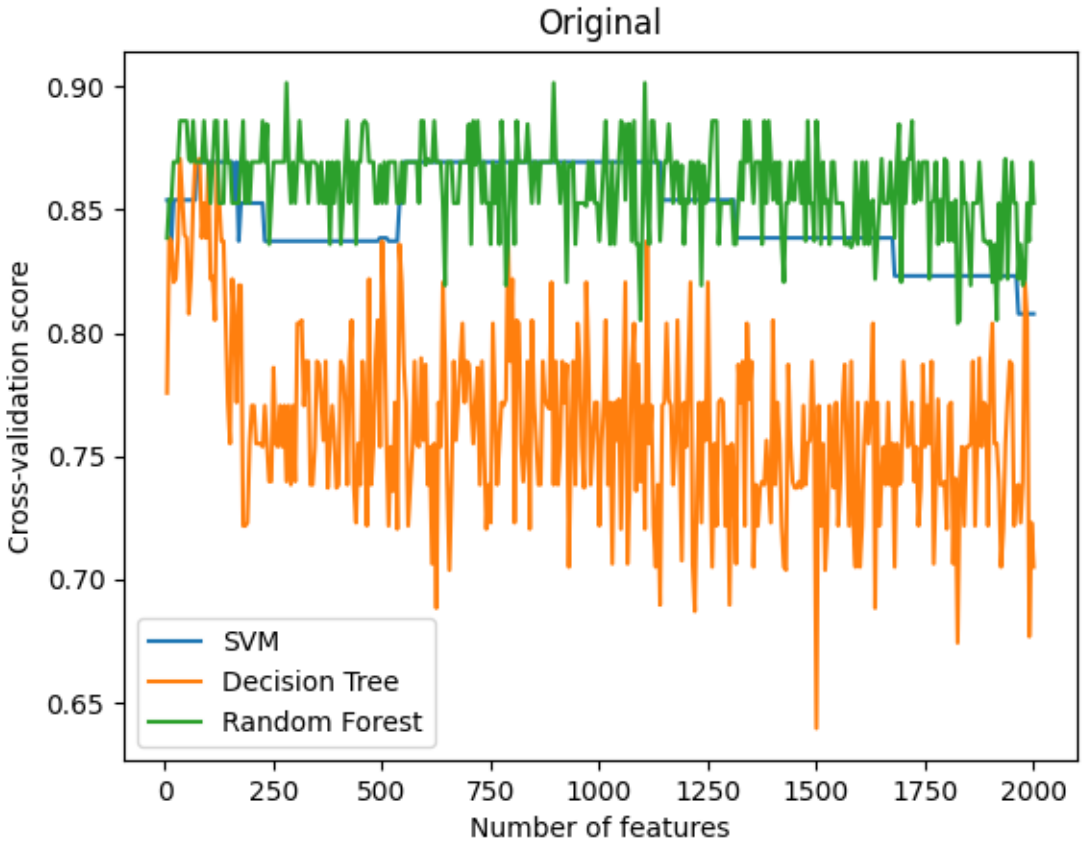
  If a feature that the feature values of samples within the same class are similar while the feature values of samples from different classes are dissimilar, then it's a good feature.

$$\text{fisher\_score}(f_i) = \frac{\sum_{j=1}^{c} n_j(\mu_{ij} - \mu_i)^2}{\sum_{j=1}^{c} n_j \sigma_{ij}^2}$$

  where c is the number of class; n_j is the number of instances in class j, μ_i is the mean of feature subset i, μ_{ij} is the mean of feature i in class j, σ_{ij}^2 is the variance of feature subset i in class j

- Results of the feature selection
  - Using SVM
    - Max of SVM: 0.8692307692307694
    - Number of features: 75
    - Selected features: Hsa.8147, Hsa.692, Hsa.37937, Hsa.1832, Hsa.692, Hsa.692, Hsa.36689, Hsa.1131, Hsa.2456, Hsa.6814, Hsa.8125, Hsa.36952, Hsa.601, Hsa.2928, Hsa.773, Hsa.957, Hsa.2344, Hsa.3306, Hsa.2097, Hsa.831, Hsa.4689, Hsa.8068, Hsa.6472, Hsa.1221, Hsa.2291, Hsa.462, Hsa.3016, Hsa.3305, Hsa.1047, Hsa.10755, Hsa.11616, Hsa.821, Hsa.5392, Hsa.1130, Hsa.1660, Hsa.11673, Hsa.2800, Hsa.5398, Hsa.1205, Hsa.4252, Hsa.14069, Hsa.1073, Hsa.43279, Hsa.2863, Hsa.33, Hsa.5444, Hsa.678, Hsa.2588, Hsa.3331, Hsa.1588, Hsa.5971, Hsa.951, Hsa.466, Hsa.878, Hsa.56, Hsa.41323, Hsa.549, Hsa.832, Hsa.1435, Hsa.490, Hsa.1902, Hsa.8010, Hsa.10664, Hsa.7395, Hsa.41260, Hsa.2645, Hsa.3152, Hsa.2705, Hsa.33965, Hsa.2451, Hsa.5346, Hsa.853, Hsa.702, Hsa.25322, Hsa.1617
  - Using Decision Tree
    - Max of Decision Tree: 0.8705128205128204
    - Number of features: 35
    - Selected features: Hsa.8147, Hsa.692, Hsa.37937, Hsa.1832, Hsa.692, Hsa.692, Hsa.36689, Hsa.1131, Hsa.2456, Hsa.6814, Hsa.8125, Hsa.36952, Hsa.601, Hsa.2928, Hsa.773, Hsa.957, Hsa.2344, Hsa.3306, Hsa.2097, Hsa.831, Hsa.4689, Hsa.8068, Hsa.6472, Hsa.1221, Hsa.2291, Hsa.462, Hsa.3016, Hsa.3305, Hsa.1047, Hsa.10755, Hsa.11616, Hsa.821, Hsa.5392, Hsa.1130, Hsa.1660

- Using Random Forest
  - Max of Random Forest: 0.9012820512820514
  - Number of features: 280
  - Selected features: Hsa.8147, Hsa.692, Hsa.37937, Hsa.1832, Hsa.692, Hsa.692, Hsa.36689, Hsa.1131, Hsa.2456, Hsa.6814, Hsa.8125, Hsa.36952, Hsa.601, Hsa.2928, Hsa.773, Hsa.957, Hsa.2344, Hsa.3306, Hsa.2097, Hsa.831, Hsa.4689, Hsa.8068, Hsa.6472, Hsa.1221, Hsa.2291, Hsa.462, Hsa.3016, Hsa.3305, Hsa.1047, Hsa.10755, Hsa.11616, Hsa.821, Hsa.5392, Hsa.1130, Hsa.1660, Hsa.11673, Hsa.2800, Hsa.5398, Hsa.1205, Hsa.4252, Hsa.14069, Hsa.1073, Hsa.43279, Hsa.2863, Hsa.33, Hsa.5444, Hsa.678, Hsa.2588, Hsa.3331, Hsa.1588, Hsa.5971, Hsa.951, Hsa.466, Hsa.878, Hsa.56, Hsa.41323, Hsa.549, Hsa.832, Hsa.1435, Hsa.490, Hsa.1902, Hsa.8010, Hsa.10664, Hsa.7395, Hsa.41260, Hsa.2645, Hsa.3152, Hsa.2705, Hsa.33965, Hsa.2451, Hsa.5346, Hsa.853, Hsa.702, Hsa.25322, Hsa.1617, Hsa.27686, Hsa.1207, Hsa.3001, Hsa.539, Hsa.1387, Hsa.31630, Hsa.9972, Hsa.2250, Hsa.1763, Hsa.2644, Hsa.9218, Hsa.1985, Hsa.43252, Hsa.1410, Hsa.2553, Hsa.1095, Hsa.1726, Hsa.3263, Hsa.41283, Hsa.544, Hsa.2715, Hsa.41280, Hsa.7, Hsa.103, Hsa.37541, Hsa.662, Hsa.954, Hsa.5211, Hsa.9353, Hsa.2196, Hsa.41280, Hsa.404, Hsa.3007, Hsa.39753, Hsa.28939, Hsa.579, Hsa.612, Hsa.28914, Hsa.16296, Hsa.789, Hsa.22762, Hsa.538, Hsa.6458, Hsa.1454, Hsa.127, Hsa.2818, Hsa.1198, Hsa.286, Hsa.26528, Hsa.36696, Hsa.229, Hsa.451, Hsa.41338, Hsa.330, Hsa.489, Hsa.558, Hsa.812, Hsa.3348, Hsa.3803, Hsa.8040, Hsa.1591, Hsa.1132, Hsa.2665, Hsa.2827, Hsa.726, Hsa.2959, Hsa.285, Hsa.8781, Hsa.1579, Hsa.41282, Hsa.929, Hsa.94, Hsa.3349, Hsa.3088, Hsa.1143, Hsa.24582, Hsa.33572, Hsa.3254, Hsa.2773, Hsa.2747, Hsa.42625, Hsa.2821, Hsa.23824, Hsa.902, Hsa.2361, Hsa.852, Hsa.2316, Hsa.37553, Hsa.1171, Hsa.36655, Hsa.627, Hsa.3566, Hsa.7652, Hsa.2753, Hsa.5544, Hsa.816, Hsa.4996, Hsa.2867, Hsa.2933, Hsa.33268, Hsa.60, Hsa.168, Hsa.2688, Hsa.2795, Hsa.21562, Hsa.3141, Hsa.305, Hsa.1592, Hsa.879, Hsa.1013, Hsa.8583, Hsa.1610, Hsa.2957, Hsa.996, Hsa.8223, Hsa.9816, Hsa.3083, Hsa.1140, Hsa.3068, Hsa.5363, Hsa.8219, Hsa.2862, Hsa.40063, Hsa.3296, Hsa.891, Hsa.7203, UMGAP, UMGAP, UMGAP, UMGAP, Hsa.3648, Hsa.39141, Hsa.2967, Hsa.5464, Hsa.491, Hsa.823, Hsa.2091, Hsa.72, Hsa.17564, Hsa.3026, Hsa.2950, Hsa.636, Hsa.1045, Hsa.698, Hsa.27685, Hsa.1978, Hsa.1121, Hsa.1013, Hsa.36694, Hsa.17901, Hsa.2221, Hsa.7048, Hsa.19731, Hsa.42949, Hsa.467, Hsa.2513, Hsa.27537, Hsa.2856, Hsa.3250, Hsa.80, Hsa.733, Hsa.2126, Hsa.1116, Hsa.2051, Hsa.6288, Hsa.36665, Hsa.35496, Hsa.12149, Hsa.612, Hsa.1689, Hsa.7462, Hsa.717, Hsa.477, Hsa.3230, Hsa.750, Hsa.2837, Hsa.41126, Hsa.3003, Hsa.9994, Hsa.61, Hsa.3002, Hsa.1994, Hsa.1567, Hsa.1714, Hsa.1254, Hsa.1043, Hsa.2917, Hsa.8126, Hsa.3963, Hsa.3093, Hsa.594, Hsa.1806, Hsa.11572, Hsa.8658, Hsa.10706, Hsa.1854, Hsa.42357, Hsa.1423, Hsa.1145, Hsa.778, Hsa.1402, Hsa.18790, Hsa.12260, Hsa.2966, Hsa.612



Problem 2 Subset-Based Feature Selection

- Metaheuristic: GA (genetic algorithm)

  Genetic Algorithms aim to replicate the behavior of genetic evolution, whereby the genetics of the individuals best suited to the environment persist over time. The individuals in the population are evaluated according to an objective function, which is used to choose the individuals to reproduce in each iteration. Those individuals that obtain a better result from the objective function will be chosen with a higher probability for reproduction. After several generations, the best individual is selected as the final result.

  - Representation of individuals:

    Individuals can be represented by different encoding ways. In this implementation, the individuals are represented as a string of length equal to the number of features, where each character of the string corresponds to a feature and has a value of 1 or 0 depending on whether the feature is active or not.

  - Starting point:

    The algorithm starts with a randomly generated population with a certain number of individuals, the number is set to be 200 in this implementation.

  - Objective function:

$F = w * c(x) + (1 - w) * (1/s(x))$

* x: a feature subset (an individual)
* w: a parameter between 0 and 1

  w is set to be 1 in this implementation.
* c(x): classification accuracy of x

  The classifier used in this implementation is decision tree.
* s(x): weighted size of the feature subset represented by x

– Termination: The iterative process terminates when the number of generations is reached. The number of generations is set to be 5 in this implementation.

– Genetic operations:

* Selection:

  The probability that an individual will be selected is proportional to its own fitness and is inversely proportional to the fitness of the other competing hypothesis in the current population.

  The fitness is the value of the objective function in the optimization problem.

  $P(x) = \frac{F(x)}{\sum_{i=1}^{P} F(x_i)}$
* Crossover:

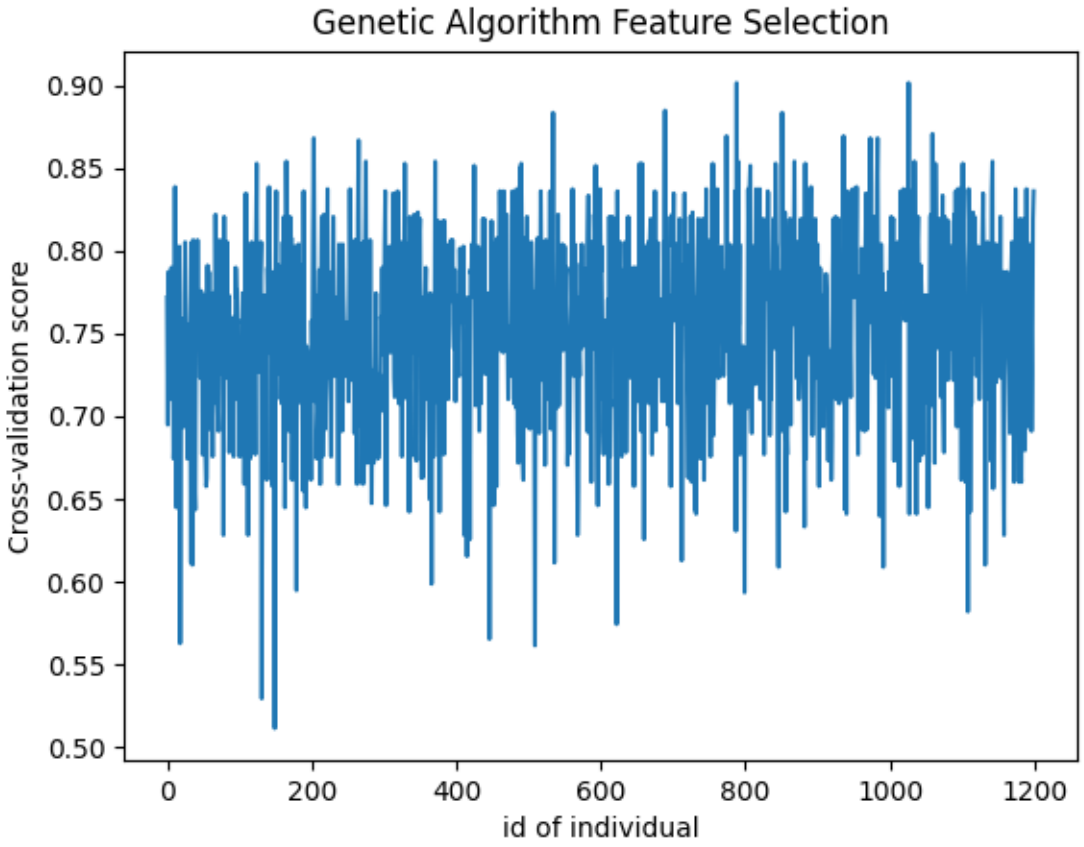
  Use single-point crossover. A number between 1 and n-1 will be randomly generated (n is the number of features). This number will be used to divide the genes of the parents into randomly sized chunks. The first of the new individuals will have the first chunk from parent A and the second chunk from parent B. The second new individual will consist of the first chunk of parent B and the second chunk of parent A.
* Mutation:

  Each individual has a probability p_m to mutate. These mutations help to create slight variations in the individuals that avoid generating populations of identical individuals. In this implementation, p_m is set to be 0.05, which means approximately 5 mutations will occur per 100 chars.

– Results of the feature selection

* Max of GA: 0.9012820512820513
* Number of features: 388
* Selected features: Hsa.13491, Hsa.541, Hsa.20836, Hsa.1977, Hsa.3087, Hsa.750, Hsa.2555, Hsa.2357, Hsa.3002, Hsa.44472, Hsa.4689, Hsa.1648, Hsa.1836, Hsa.1273, Hsa.2948, Hsa.7877, Hsa.1400, HSAC07, HSAC07, HSAC07, HSAC07, Hsa.8068, Hsa.1139, Hsa.3006, Hsa.2299, UMGAP, UMGAP, Hsa.1238, Hsa.98, Hsa.8093, Hsa.5444, Hsa.361, Hsa.8125, Hsa.2361, Hsa.255, Hsa.2699, Hsa.44403, Hsa.9304, Hsa.2063, Hsa.36957, Hsa.33965, Hsa.2219, Hsa.2665, Hsa.2151, Hsa.1500, Hsa.45222, Hsa.2419, Hsa.1985, Hsa.1978, Hsa.5346, Hsa.4316, Hsa.914, Hsa.7280, Hsa.1737, Hsa.1939, Hsa.2929, Hsa.21232, Hsa.13183,

Hsa.572, Hsa.45260, Hsa.3225, Hsa.6977, Hsa.41309, Hsa.957, Hsa.2071, Hsa.91, Hsa.558, Hsa.1222, Hsa.652, Hsa.1367, Hsa.63, Hsa.3197, Hsa.22614, Hsa.36957, Hsa.10306, Hsa.733, Hsa.2357, Hsa.24539, Hsa.45604, Hsa.19, Hsa.21418, Hsa.18664, Hsa.10510, Hsa.514, Hsa.1137, Hsa.2831, Hsa.2015, Hsa.2244, Hsa.832, Hsa.2773, Hsa.1822, Hsa.2463, Hsa.3003, Hsa.812, Hsa.8305, Hsa.2360, Hsa.570, Hsa.304, Hsa.2846, Hsa.915, Hsa.546, Hsa.991, Hsa.31882, Hsa.587, Hsa.2095, Hsa.3295, Hsa.4954, Hsa.4992, Hsa.1006, Hsa.21339, Hsa.8656, Hsa.9623, Hsa.692, Hsa.8177, Hsa.8147, Hsa.41315, Hsa.996, Hsa.33572, Hsa.692, Hsa.1825, Hsa.81, Hsa.5122, Hsa.539, Hsa.10358, Hsa.3769, Hsa.80, Hsa.29228, Hsa.20034, Hsa.43279, Hsa.1050, Hsa.9407, Hsa.21660, Hsa.9720, Hsa.3063, Hsa.2829, Hsa.17256, Hsa.1178, Hsa.26767, Hsa.1534, Hsa.1166, Hsa.2779, Hsa.1722, Hsa.122, Hsa.2768, Hsa.3272, Hsa.1263, Hsa.2902, Hsa.1063, Hsa.3121, Hsa.1044, Hsa.2917, Hsa.13598, Hsa.1391, Hsa.698, Hsa.1738, Hsa.14360, Hsa.9738, Hsa.17649, Hsa.2598, Hsa.1276, Hsa.2126, Hsa.2559, Hsa.1466, Hsa.35496, Hsa.31, Hsa.153, Hsa.3910, Hsa.194, Hsa.1277, Hsa.4347, Hsa.13702, Hsa.13007, Hsa.758, Hsa.7671, Hsa.2943, Hsa.2440, Hsa.20507, Hsa.6160, Hsa.1687, Hsa.27832, Hsa.1008, Hsa.1278, Hsa.45242, Hsa.1121, Hsa.5971, Hsa.1558, Hsa.142, Hsa.28145, Hsa.16100, Hsa.5537, Hsa.9667, Hsa.8097, Hsa.41282, Hsa.19731, Hsa.244, Hsa.3328, Hsa.551, Hsa.1896, Hsa.2119, Hsa.3185, Hsa.50, Hsa.9285, Hsa.2951, Hsa.7247, a.1000, Hsa.2589, Hsa.26083, Hsa.2191, Hsa.1591, Hsa.962, Hsa.1312, Hsa.268, Hsa.23285, Hsa.3209, Hsa.8964, Hsa.2487, Hsa.17822, Hsa.847, Hsa.3135, Hsa.318, Hsa.26883, Hsa.2840, Hsa.978, Hsa.2179, Hsa.44686, Hsa.34384, Hsa.2856, Hsa.9235, Hsa.2135, Hsa.3298, Hsa.773, Hsa.3353, Hsa.316, Hsa.3145, Hsa.25, Hsa.36685, Hsa.2105, Hsa.2955, Hsa.20524, Hsa.35518, Hsa.1219, Hsa.3154, Hsa.45754, Hsa.14884, Hsa.5532, Hsa.1294, Hsa.17595, Hsa.742, Hsa.8096, Hsa.12102, Hsa.2565, Hsa.1598, Hsa.1798, Hsa.10706, Hsa.5756, Hsa.2841, Hsa.22909, Hsa.45732, Hsa.4937, Hsa.7646, Hsa.7802, Hsa.114, Hsa.2157, Hsa.549, Hsa.9856, Hsa.40009, Hsa.1866, Hsa.1331, Hsa.440, Hsa.94, Hsa.25481, Hsa.8833, Hsa.1047, Hsa.27085, Hsa.3118, Hsa.816, Hsa.1171, Hsa.10308, Hsa.7203, Hsa.1994, Hsa.3269, Hsa.1670, Hsa.2339, Hsa.8175, Hsa.399, Hsa.290, Hsa.3271, Hsa.1383, Hsa.44676, Hsa.37931, Hsa.24867, Hsa.2774, Hsa.981, Hsa.2549, Hsa.1479, Hsa.1517, Hsa.43894, Hsa.1460, Hsa.3026, Hsa.6786, Hsa.2410, Hsa.7462, Hsa.814, Hsa.712, Hsa.16622, Hsa.3301, Hsa.36671, Hsa.2827, Hsa.449, Hsa.1775, Hsa.42746, Hsa.34510, Hsa.27300, Hsa.25522, Hsa.3263, Hsa.17210, Hsa.5514, Hsa.3071, Hsa.33867, Hsa.7741, Hsa.2426, Hsa.3127, Hsa.108, Hsa.2330, Hsa.26945, Hsa.10658, Hsa.27481, Hsa.2808, Hsa.5226, Hsa.818, Hsa.24368, Hsa.11582, Hsa.29940, Hsa.2742, Hsa.2928, Hsa.9246, Hsa.43304, Hsa.7781, Hsa.2968, Hsa.17426, Hsa.2692, Hsa.2458, Hsa.2926, Hsa.1454, Hsa.3083, Hsa.9671, Hsa.37654, Hsa.39777, Hsa.39731, Hsa.43862, Hsa.37811, Hsa.4286, Hsa.2645, Hsa.67, Hsa.3254, Hsa.230, Hsa.2842, Hsa.3324, Hsa.32907, Hsa.23600, Hsa.1811, Hsa.42255, Hsa.410, Hsa.2253, Hsa.2753, Hsa.10746, Hsa.37609, Hsa.565, Hsa.3862, Hsa.2029, Hsa.2602, Hsa.2573, Hsa.17514, Hsa.1780, Hsa.2456, Hsa.2939, Hsa.8192, Hsa.2519, Hsa.8503, Hsa.41306, Hsa.33095, Hsa.9174, Hsa.3005, Hsa.36696, Hsa.44116, Hsa.17901


Genetic Algorithm Feature Selection
Cross-validation score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
200
400
600
800
1000
1200
id of individual

Problem 3 ARIMA Forecast

- The ARIMA parameters used in my implementation: (0, 1, 0, 0, 0, 0)
- The MSE: 79345.216335413