FEATURE SELECTION: SURVEY ON SOME METHODS

1/ Approach the problem:

To be honest, the solution of the problem already is provided [1]. After reviewing the reference of that solution [2], I am aware of the pros and con of the method [2] mentions by implementing them and have some reflections (include the method [1] exploited). I would like to go through the choices that I had thought it could be possible

2/ Mean decrease impurity method:

This is what [1] picked to conduct. Its idea bases on the either gini or information(entropy) calculated when building decision tree or random forest particularly (multi-decision tree for subsampling models). While random forest itself has bias problem with variable importance measures [3], if there are correlated features, other features are underestimated if one is surveyed [2]. Because I have no information about relationship of features in data, I decided to skip this method

3/ Recursive feature elimination:

Offered by part4 of [2]: while it sounds objective to the nature of model, it changes model itself in the next iterative step by eliminating the “worst feature”, RFE conducted with random forest

4/ coefficients of regression

Part 2 of [2] offers a criteria of feature selection based on coefficients of linear model, L2 regularization / Ridge regression is the best choice to avoid linear correlated feature of model. However, I expect to have a method that can automatically iterate the steps to reach a stability over time, better than a single run.

5/ stability-selection:

Part 4 of [2] recommends a stability selection which runs several times of a selection algorithm on subsets of data with different subset of features and “score” features. This method is trusted as “makes variable selection consistent in settings where the original methods fail’ [4].

Stability selection provided in [2] is deprecated, I therefore use an open source [5]

6/ Result & Evaluation

method	(1)Mean decrease impurity	(2)Recursive feature elimination	(3)Ridge regression	(4)stability-selection
Ranking (by index of sensor)	8-6-4-0-2-7-3-1-9-5	8-6-4-0-2-1-7-5-9-3	8-4-0-3-7-9-5-2-6-1	8-4-3-0-7-1-9-5-6-2

Both methods related to decision tree (1) and (2) put 6 at top rank and 9 at the bottom. (1), (2) also give different results every shoot

I CHOOSE (40 AS SOLUTION

7/ Pros and Con of my solution

While I believe that stability selection is the best way in this approach by its transcendencies listed in [4], I did not have time to overhaul the model more. Some acts may improve the quality of this ranking:

1/ Understand more about feature relation: some calculations of co-variance and correlation coefficients can contribute to pick the best method

2/ switch into some other models of stability selection: I chose a pipeline with LogisticRegression, a randomized lasso could be tried

3/ tune some factors of stability selection: lambda_grid in example

REFERENCE

[1] https://github.com/olpotkin/CeleraOne-Solution/blob/master/Solution.ipynb

[2] http://blog.datadive.net/selecting-good-features-part-iii-random-forests/

[3] Bias in random forest variable importance measures: Illustrations, sources and a solution,BMC Bioinformatics2007, Carolin StroblEmail, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn

[4] Stability Selection, Nicolai Meinshausen and Peter B¨uhlmann,University of Oxford and ETH Z¨urich, May 16, 2009

[5] https://github.com/scikit-learn-contrib/stability-selection