Università degli Studi di Padova

Dipartimento Matematica

Robust feature selection

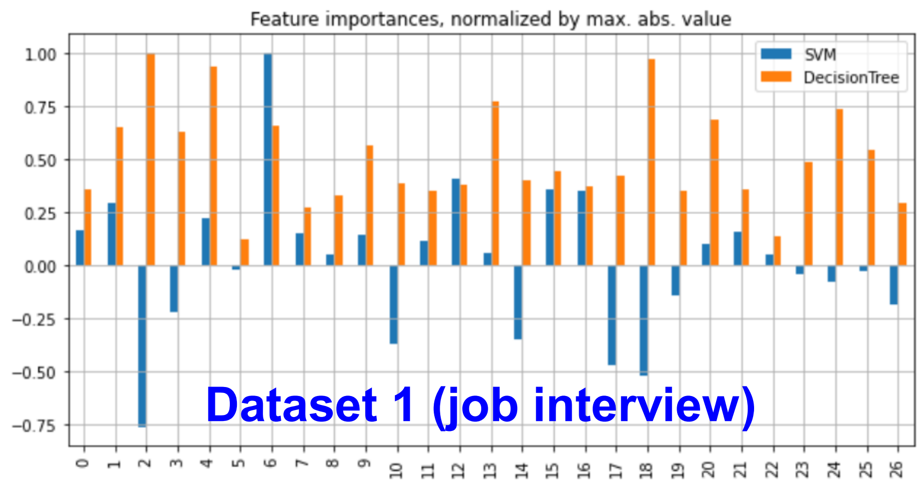
Manoj Kumar Nagabandi
Andrii Kliachkin
Pavel Ianko
pavel.ianko@studenti.unipd.it
University of Padua, Data Science MSc

Overview

- Do diverse models select different items as important?
  - Is there a difference in response to fake-good / fake-bad datasets?
  - Does result change across datasets?
- Feature selection techniques
  - Accuracy with 20% best features
  - Accuracy drop for reduced datasets
  - Concordance in spotting 20% items
- Agnostic VS model dependent feature selection
  - PCR weaknesses
  - Goals
  - Related work
- Psychometric VS model dependent feature selection
- Comments

Do diverse models select different items as important?

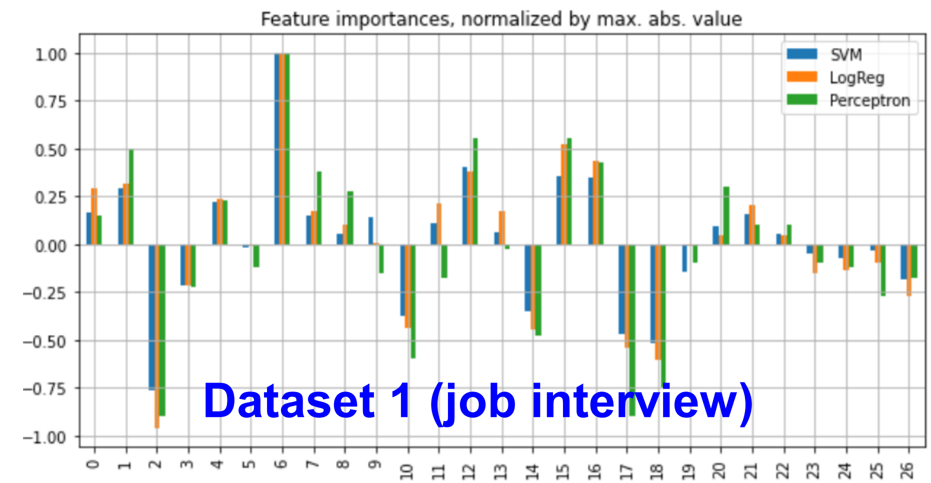
Do models select different features?



Processing pipeline:

- Ordinal encoder for the target variable
- Default sklearn hyperparameters, fixed random state
- Max-norm feature importance

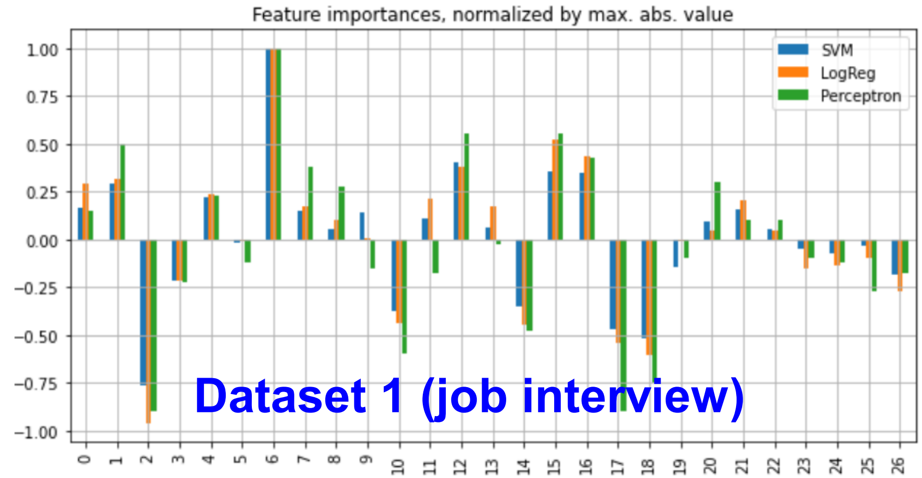
Do models select different features?

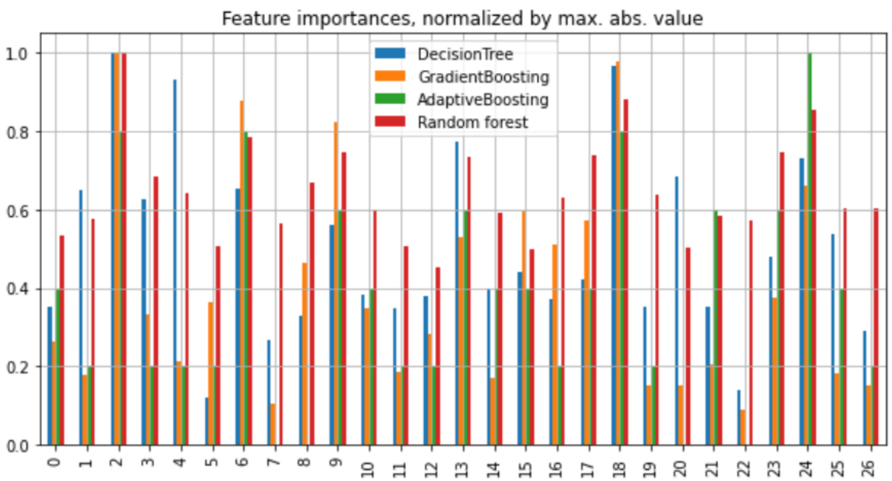


Processing pipeline:

- Ordinal encoder for the target variable
- Default sklearn hyperparameters, fixed random state
- Max-norm feature importance

Do models select different features?



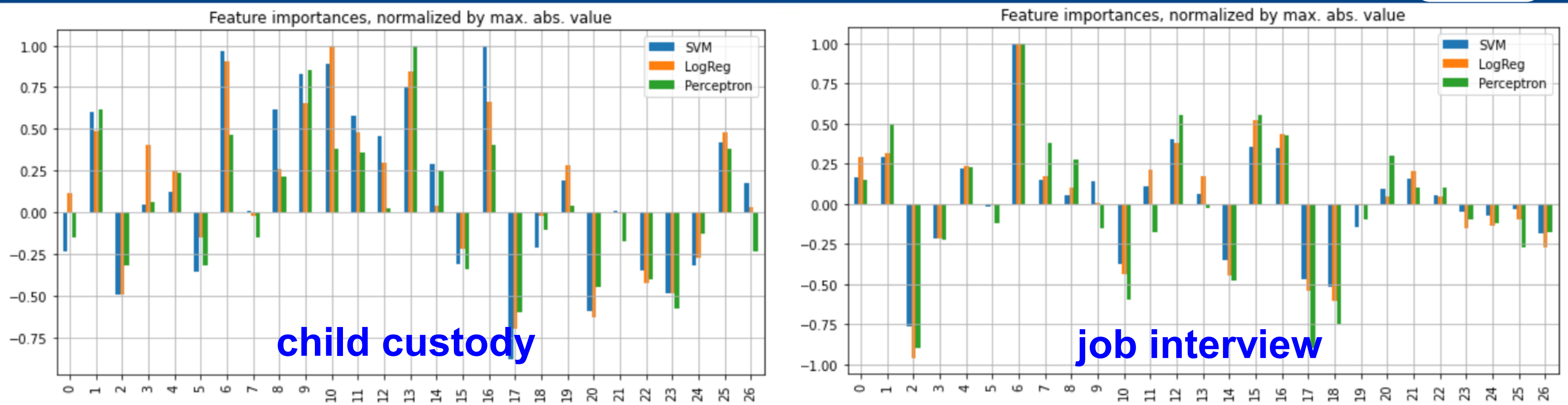


Dataset 1 (job interview)

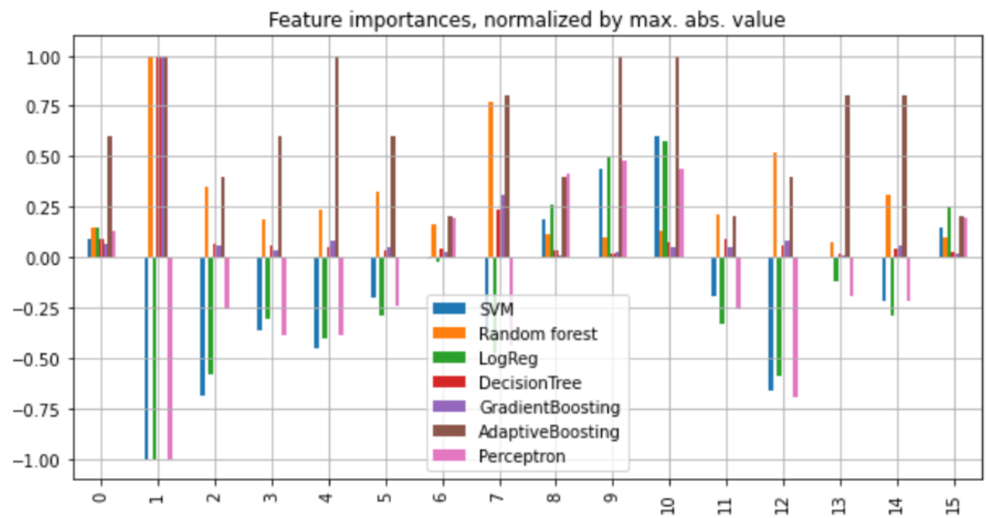
Processing pipeline:

- Ordinal encoder for the target variable
- Default sklearn hyperparameters, fixed random state
- Max-norm feature importance

Any difference to fake good / fake bad?

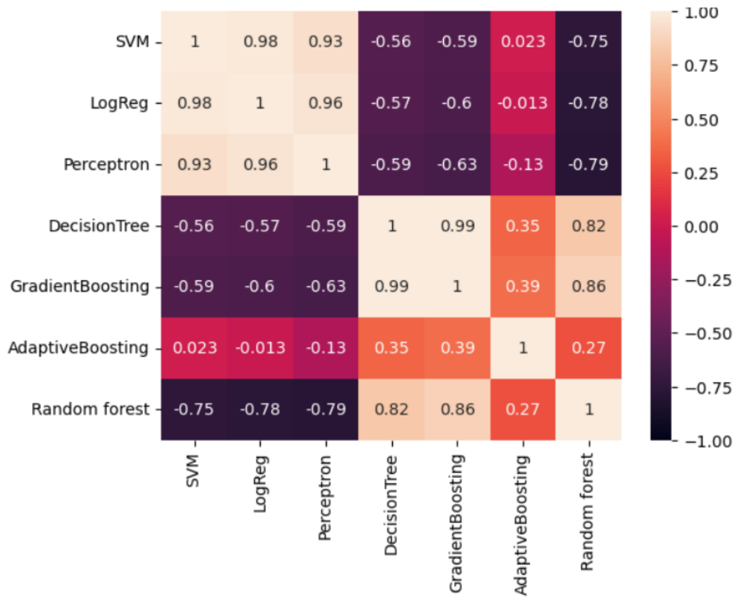


Dataset 1 (short Dart Triad, Fake Good)

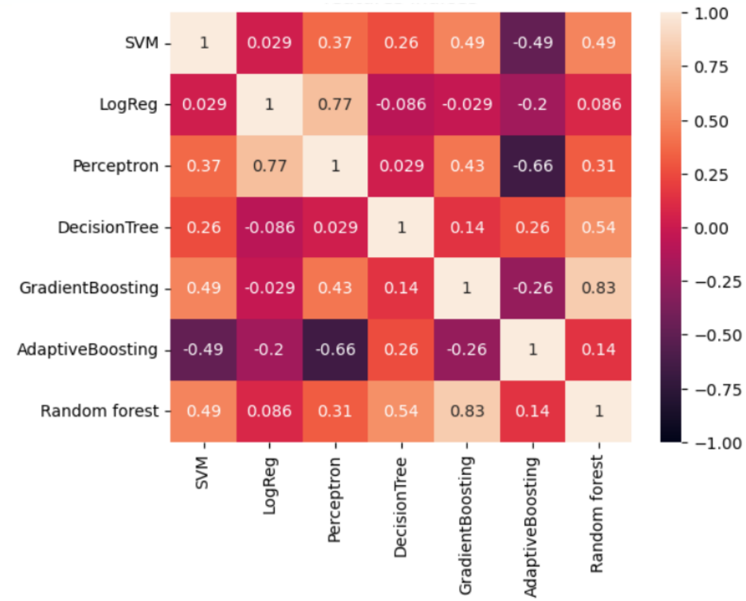


Dataset 2 (PMRQ, Fake Bad)

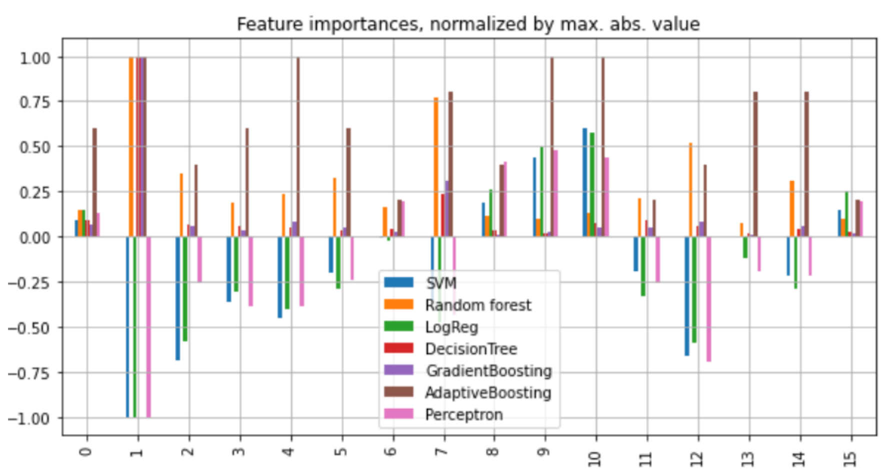
Tree-based models VS scalar product based



Feature importance correlation



Rank order correlation



Dataset 2 (PMRQ, Fake Bad)

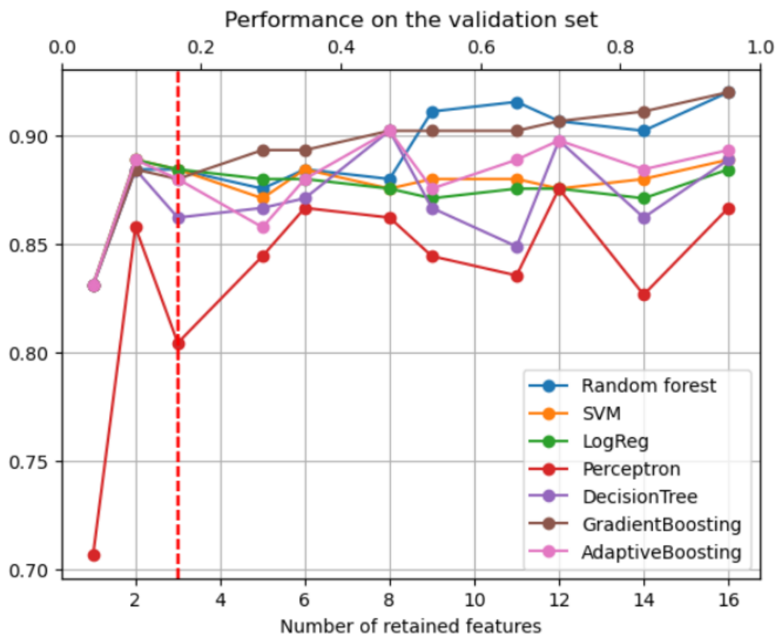
Answer:

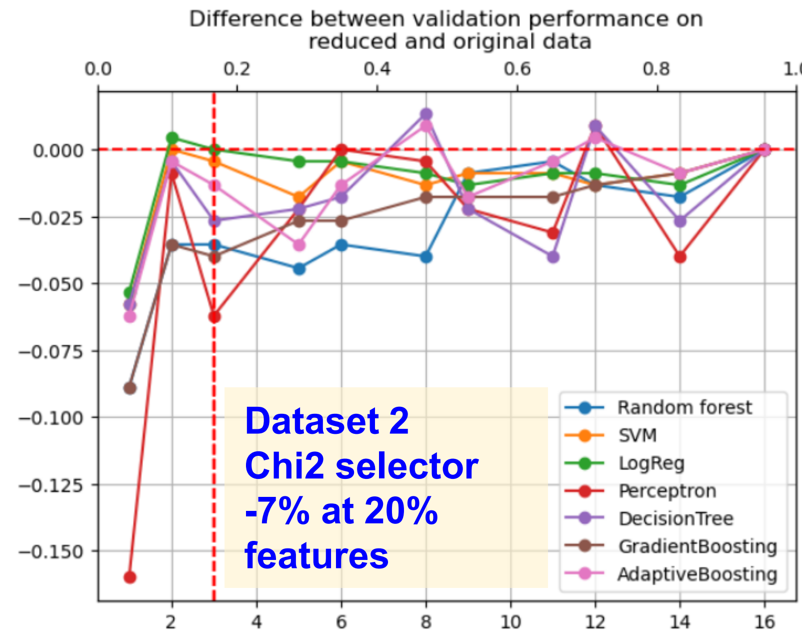
- Feature importance patterns are different for tree-based / product-based models
- Rank order correlation is weak → models select different features

Model-independent feature selection techniques

Feature selection techniques

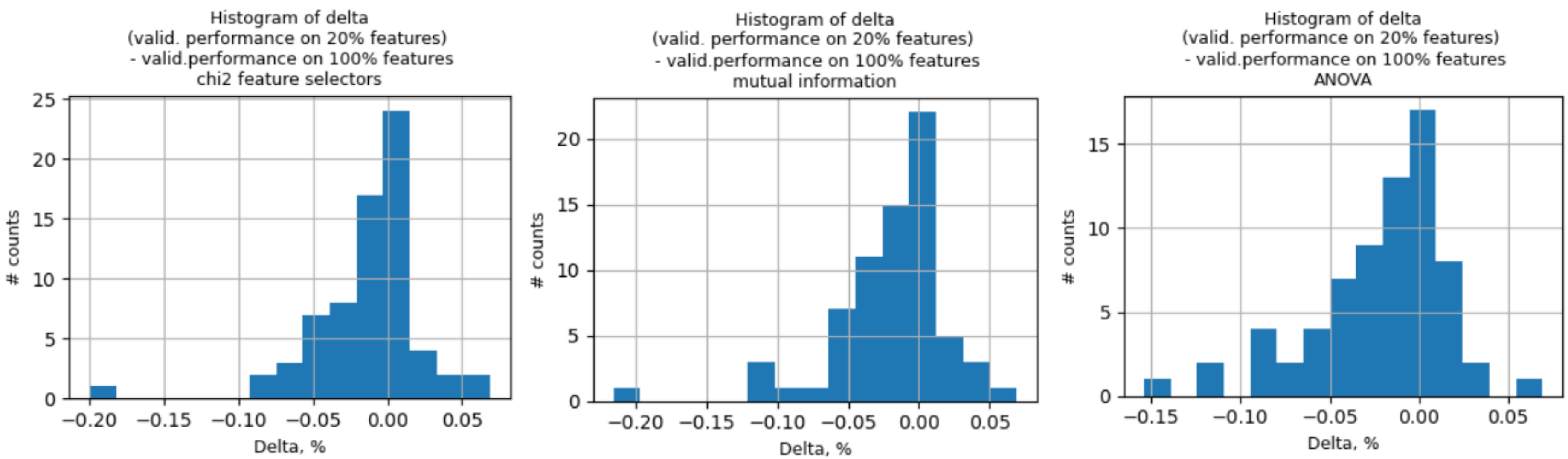
- **Model-independent and data independent** feature selection:
  - For categorical input + categorical output [1]:
    - **Chi-squared** [1]
    - **Mutual information** [1]
  - For numerical input + categorical output [1]:
    - **ANOVA testing** [1]





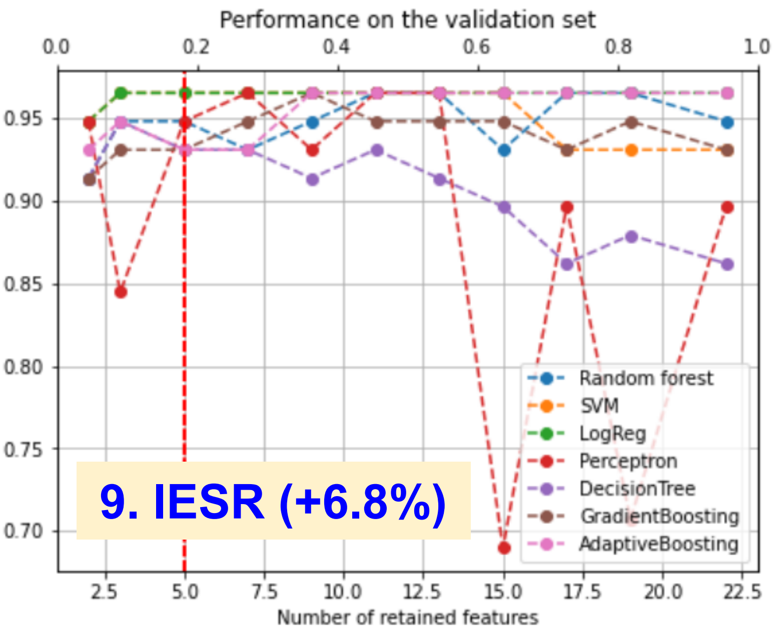
[1]. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Summarized results

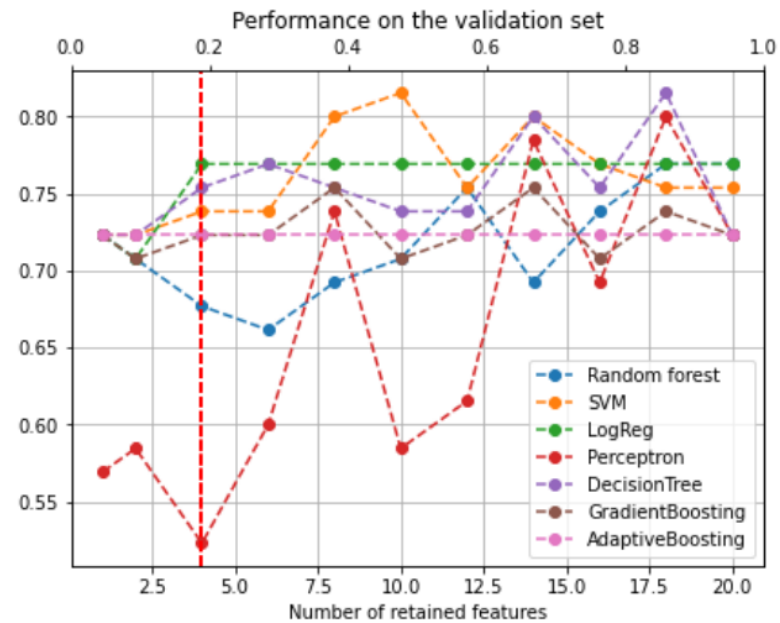


Dataset	Models	Feature selector	100% features, val. acc., %	20% features, val. acc, %	(acc_20 – acc_100), %	Comment
3. PCL	Perceptron	Mut. Inf.	72.3	50.7	-21	Worst result
9. IESR	Decision tree	chi2	86.2	93.1	+6.8	Best result
9. IESR	Perceptron	Mut. Inf.	89.6	96.5	+6.8	Best result
9. IESR	Perceptron	ANOVA	89.6	96.5	+6.8	Best result

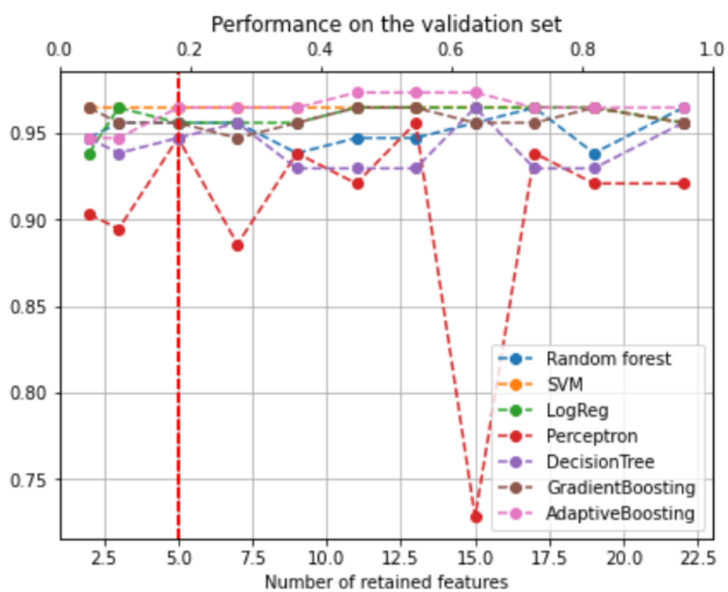
Summarized results



9. IESR (+6.8%)



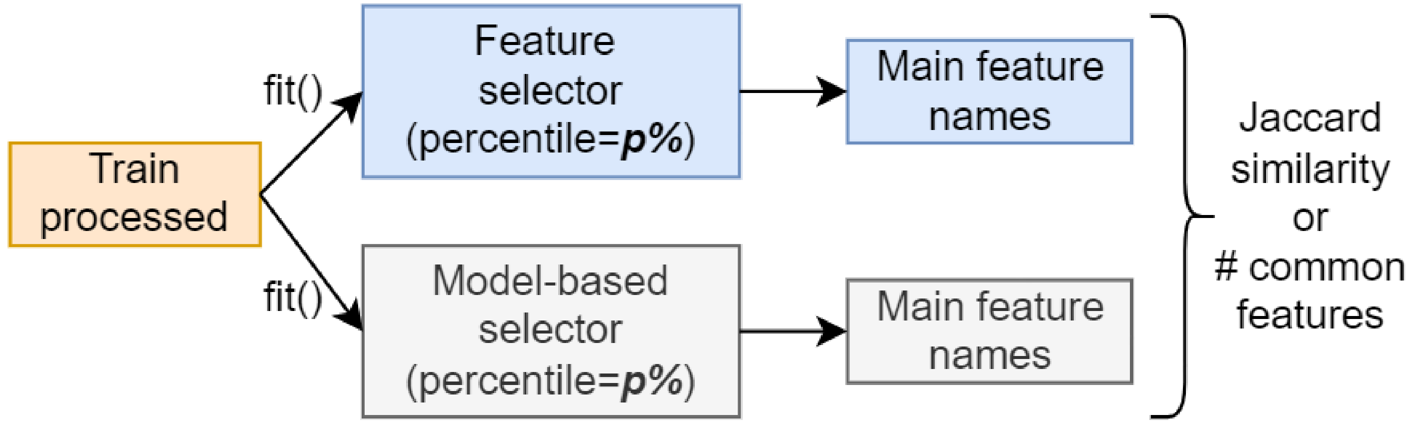
3. PCL (-20%)

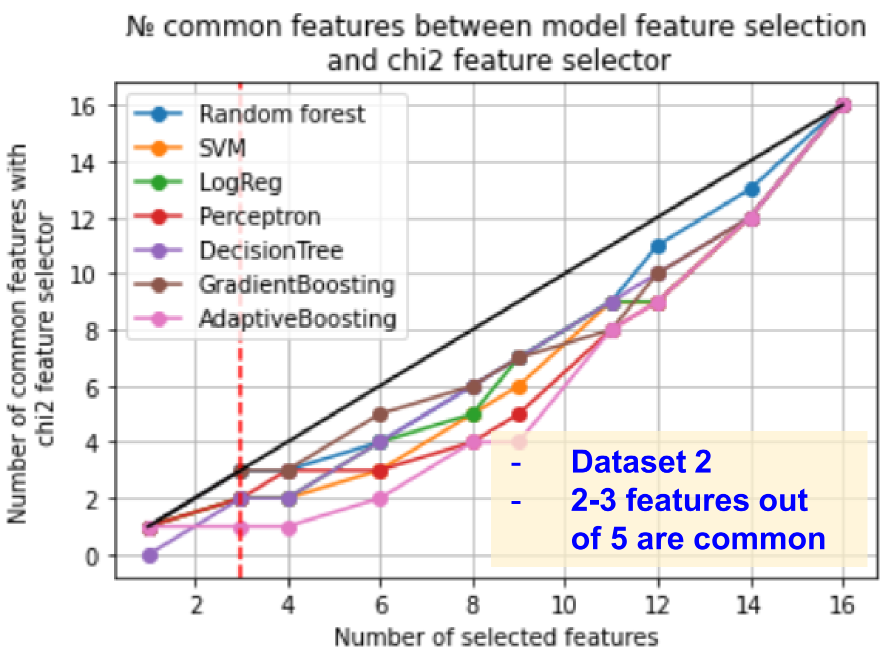


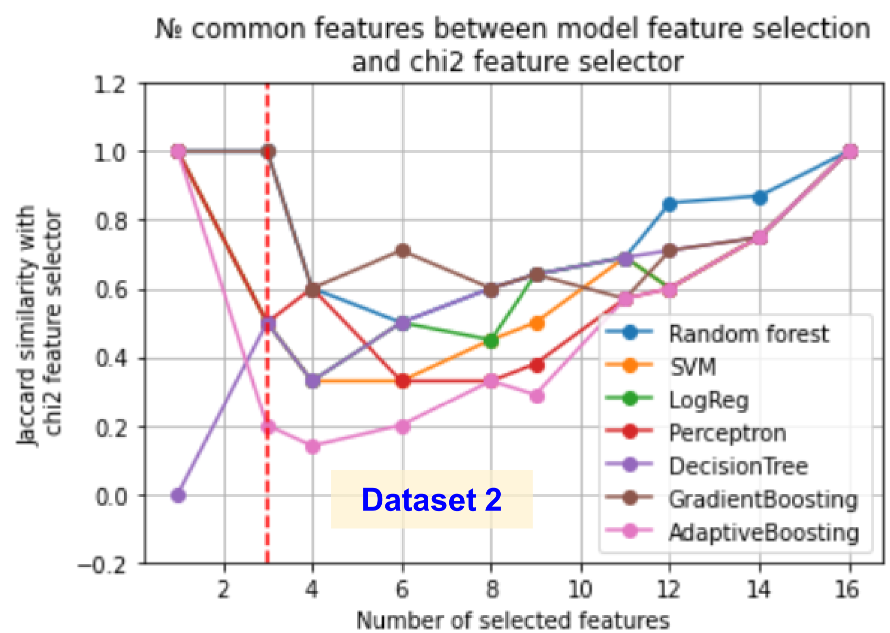
4. NAQ_R (~0%)

Concordance in spotting 20% items

∀percentile,
∀feature selector,
∀model



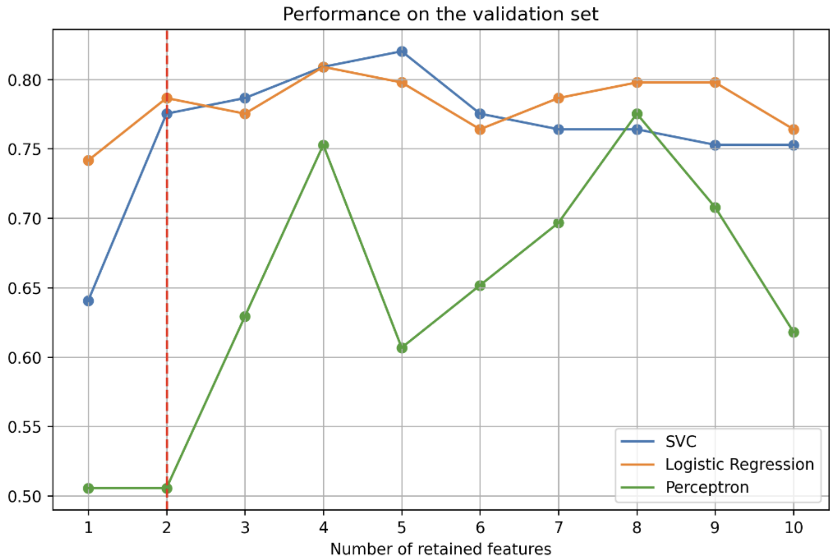


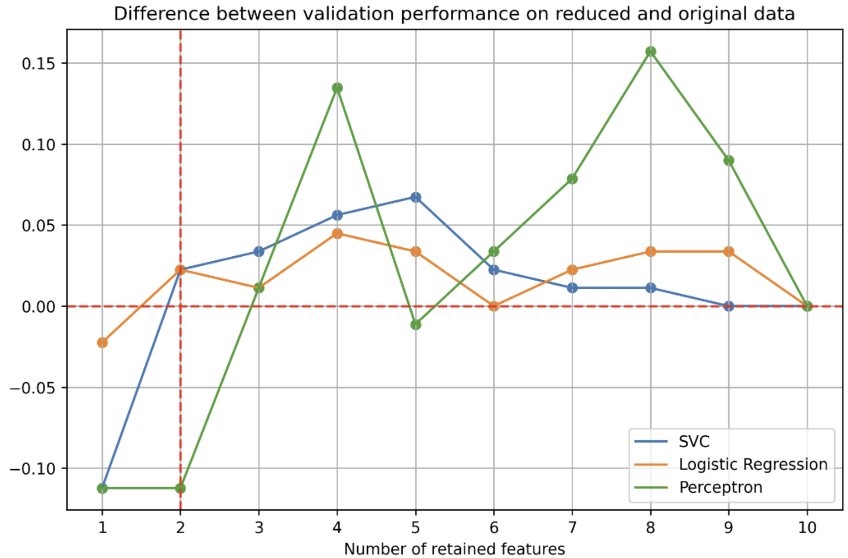


Model-dependent feature selection techniques

Model-dependent feature selection

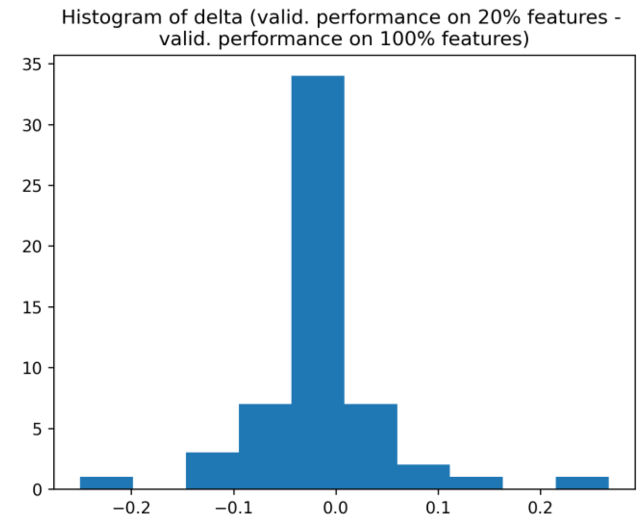
- **Model-dependent** feature selection:
  - Recursive feature elimination
- **Procedure:**
  - Model trained on the training set with full *k* features
  - Feature importance calculated from the model
  - Least important feature eliminated
  - Model retrained on *k-1* features





Model-dependent feature selection

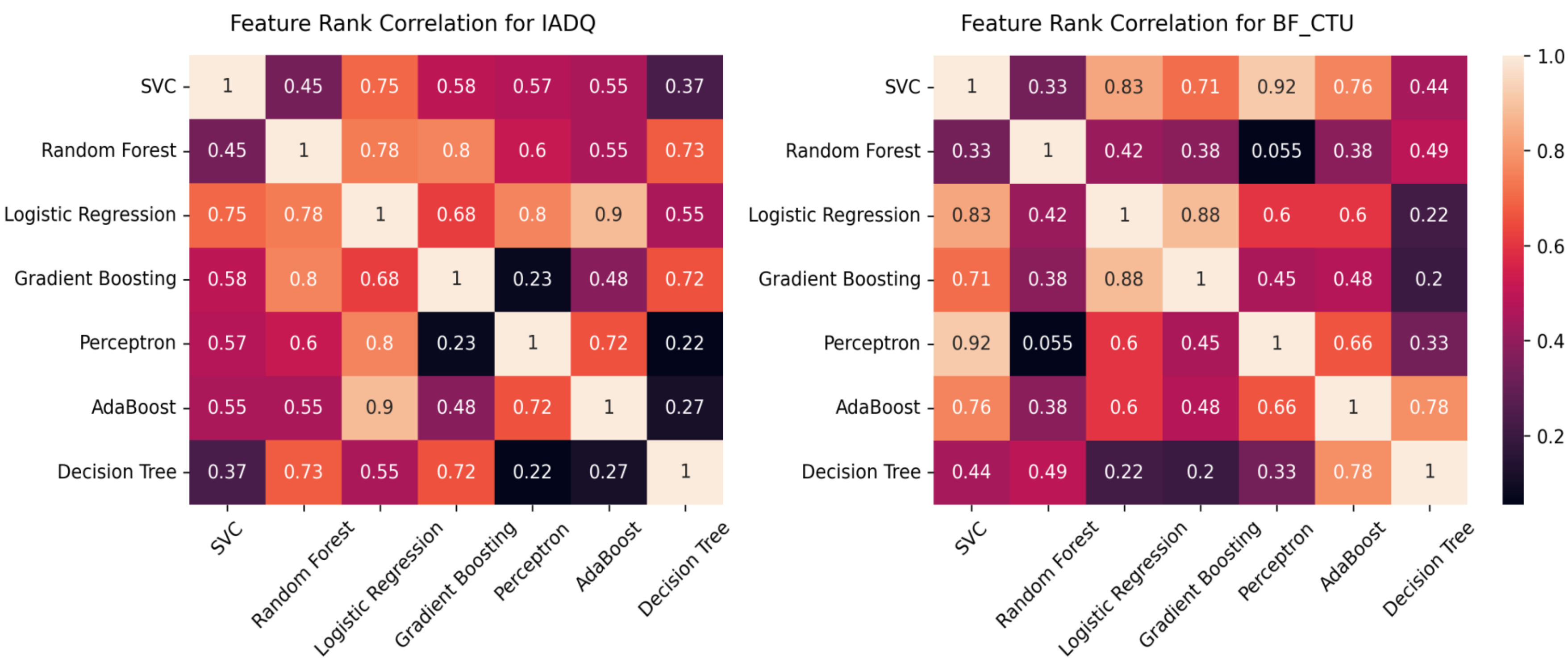
- Summarized results across 7 models:
  - Logistic regression
  - Perceptron
  - SVM
  - Random forest
  - Decision tree
  - Gradient boosting
  - Adaptive boosting



Dataset	Models	100% features, validation accuracy, %	20% features, validation accuracy, %	Delta (acc_20 - acc_100), %	Comments
3. PCL	Perceptron	75.3	62.9	-12.3	Worst drop
12. IADQ	Perceptron	50	76.6	+26.6	Best increase
5. PHQ9_GAD7	Logistic Regression	99.1	97.7	-1.3	Best 20% validation accuracy

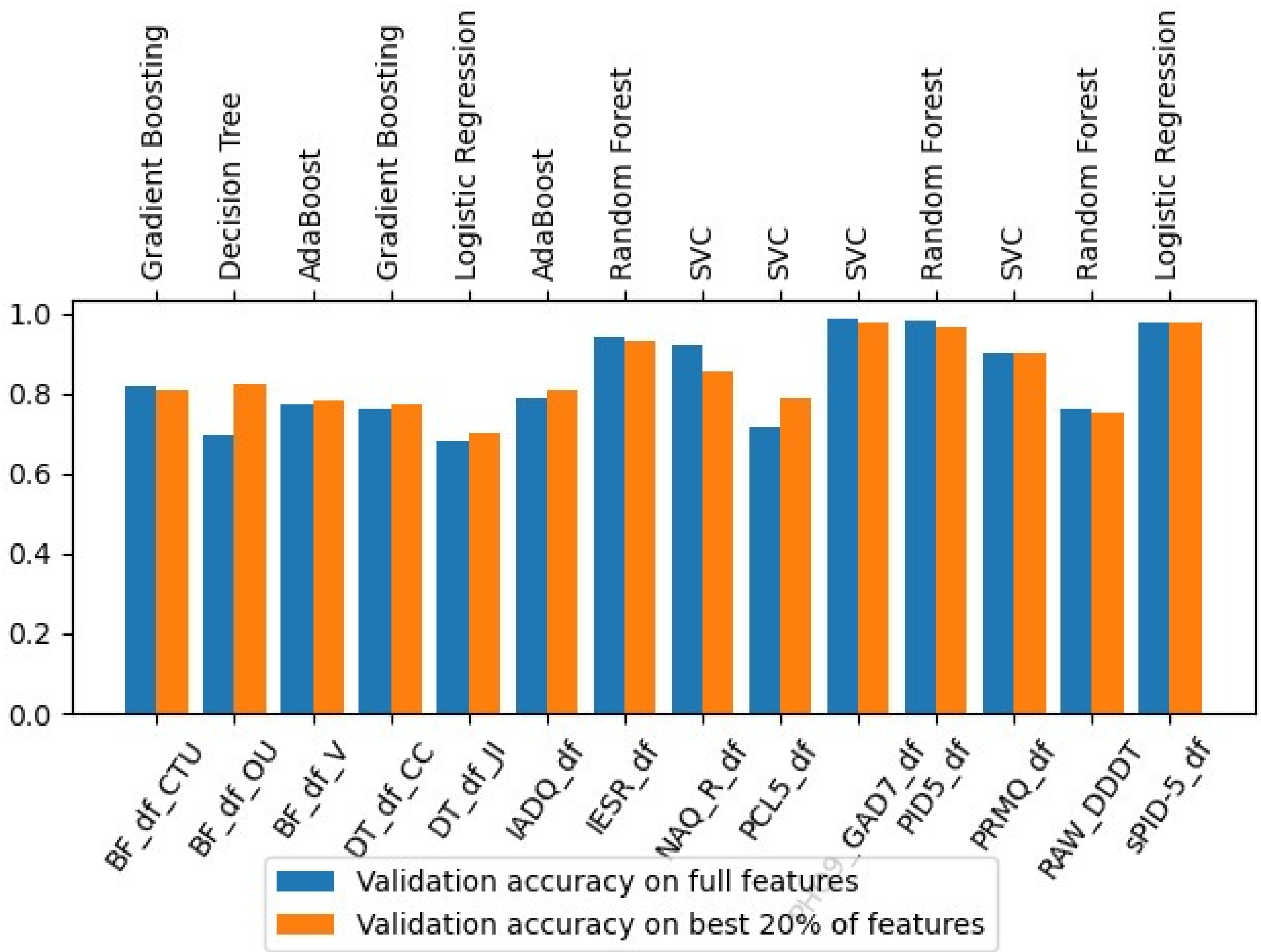
Similarity of chosen features

RFE offers low stability: different models often select different features for the same data



Model-dependent feature selection

- Summary across 7 models and 12 datasets
- Best models for the datasets are below:



Feature selection with PCA and FA

Feature selection techniques for PCA

PCA explains total variance among variables and chooses components as linear combination of variables which accounts for the max. Variance.

From each dataset only Honest Reviews are considered for PCA feature selection

Consider no. of Principal Components = total features od dataset for performing PCA

We consider topmost feature in each component based on highest explained variance in that component.

After applying feature selection technique using PCA take the top 20 % and 100 % features and compare performance of models.

Feature selection techniques for FA

Factor Analysis is a useful approach to find latent variables which are not directly measured in a single variable but rather inferred from other variables in the dataset.

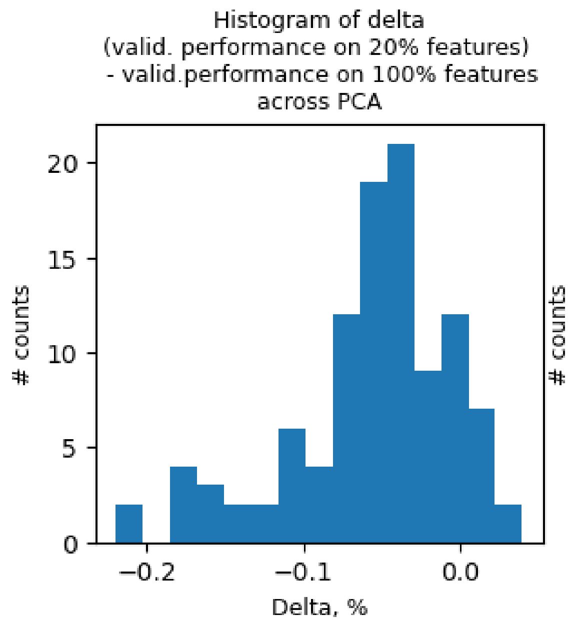
From each dataset only Honest Reviews are considered for FA feature selection.

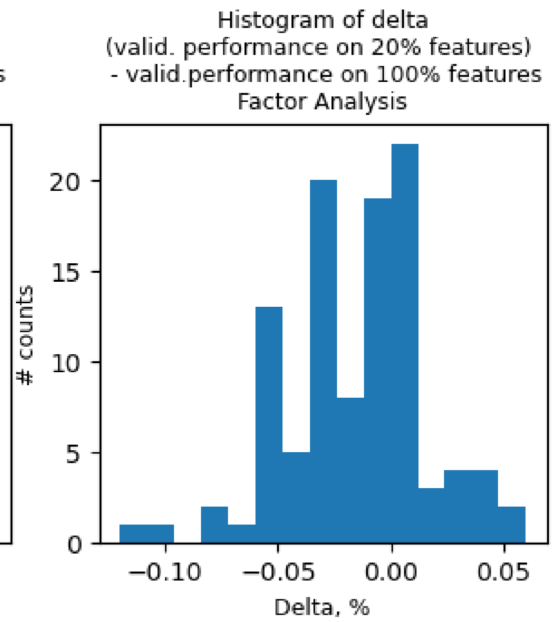
Consider no. of Factors = total features in the dataset for performing Factor Analysis

From each factor select one dominant feature based on highest loading.

After applying feature selection technique using FA take the top 20 % and 100 % features and compare performance of models.

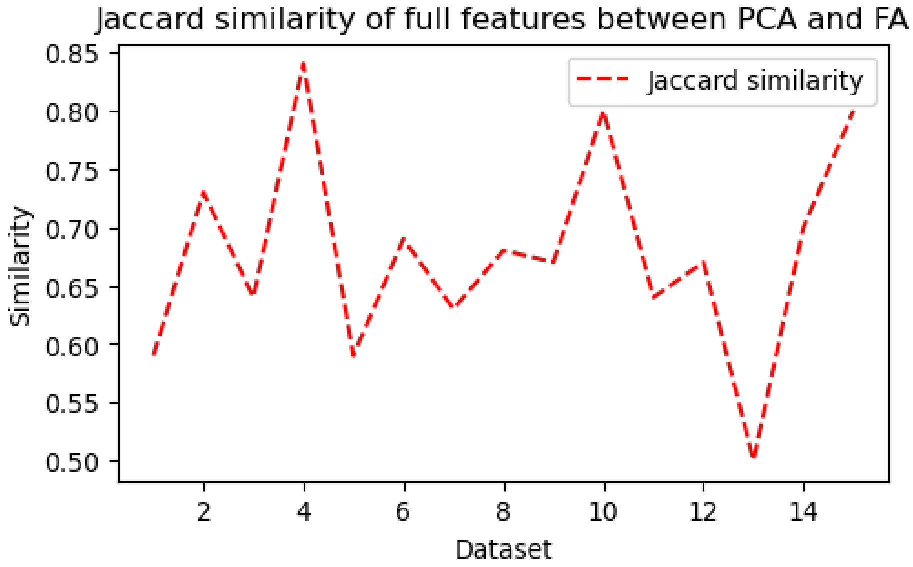
Summarized results

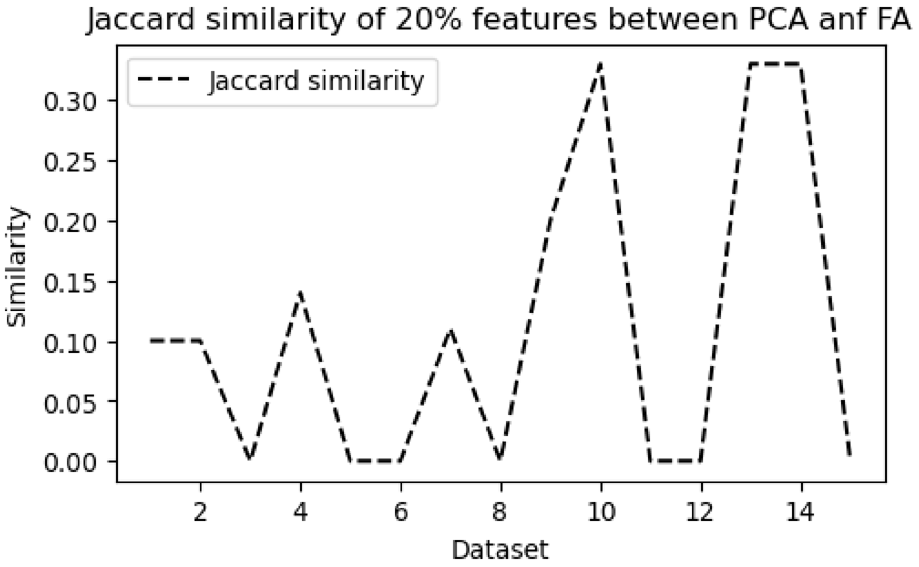




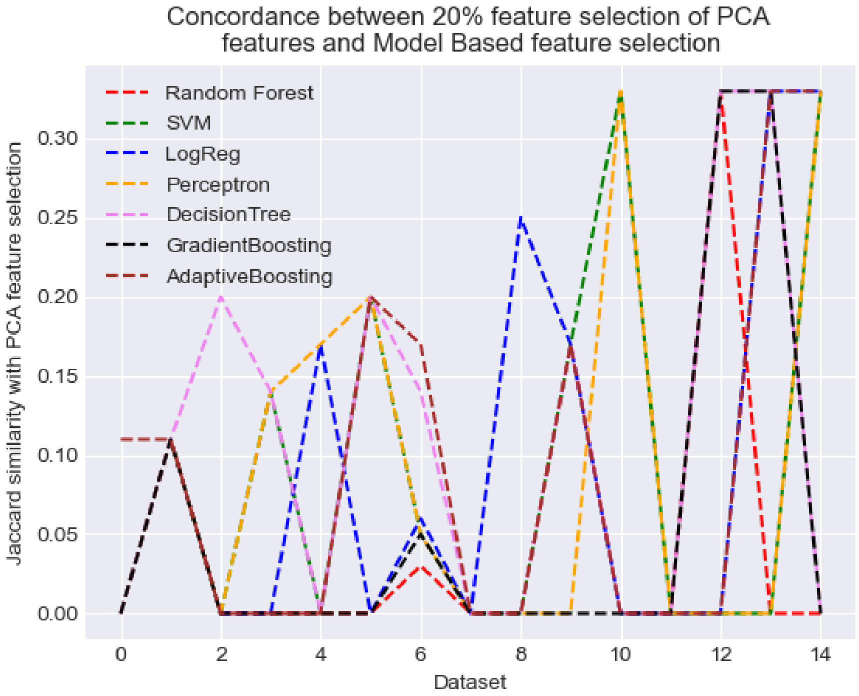
Dataset	Models	Feature selector	100% features, val. acc., %	20% features, val. acc, %	(acc_20 – acc_100), %	Comment
1. SHORTDT (cc)	Perceptron	PCA	85	63	-22	Worst result
13.BF(3)(v)	Perceptron	FA	64	70	+6	Best result
4. NAQ_R	Decision tree	FA	90	95	+5	Best result
9. IESR	Decision tree	PCA	84	88	+4	Best result

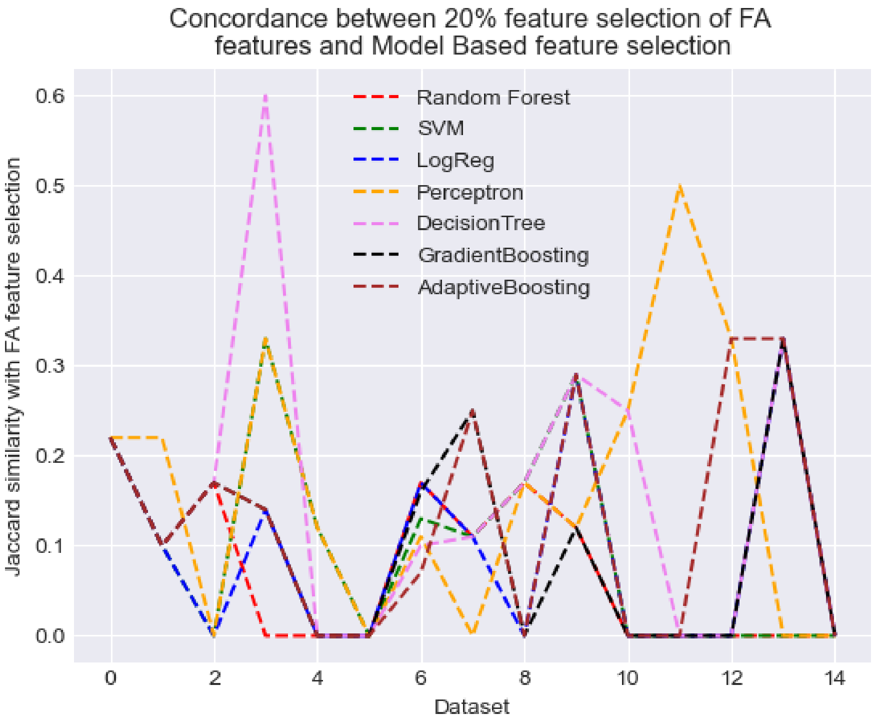
Jaccard Similarity between PCA and FA





Jaccard similarity between 20% features

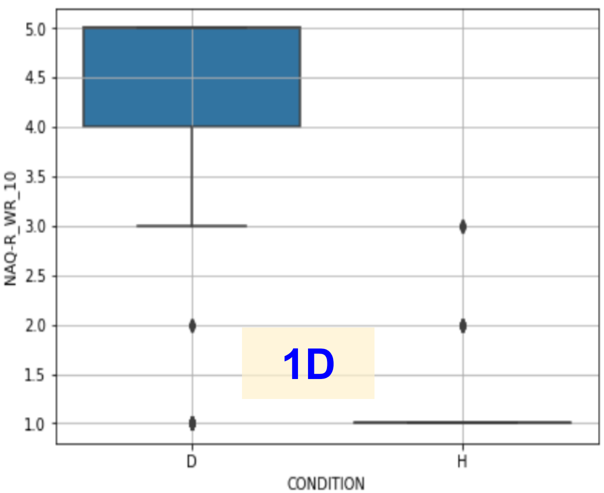




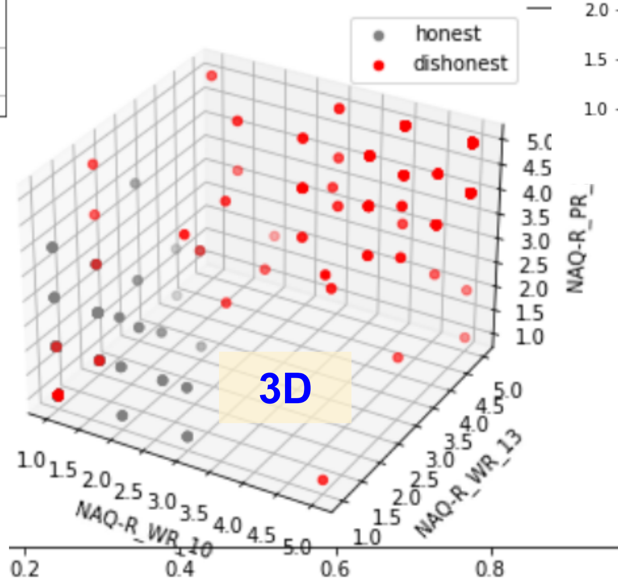
Comments

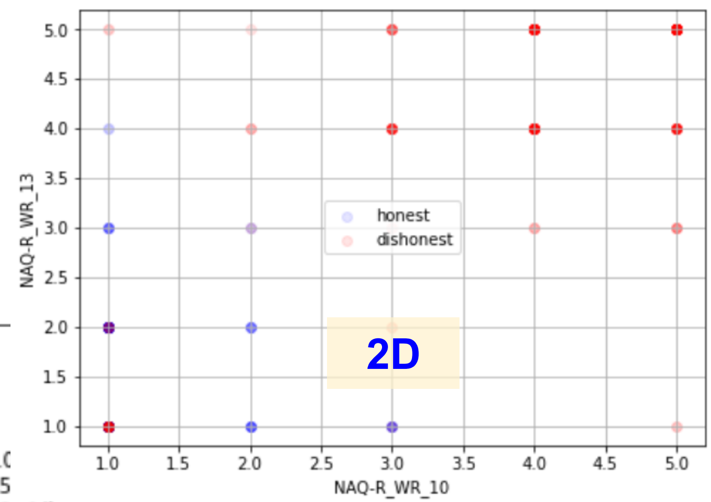
Clustering in 1D, 2D and 3D

- We can use three most important items as original (not processed!) clustering dimensionalities
- Using original items, we do not lose interpretability, as in PCA
- However, clusters are not distinct everywhere



Dataset 4





Conclusions

- Most important features as **clustering dimensionalities**
- **Tree-based** models VS **scalar product-based** models
- **3 model-independent techniques**:
  - Highest accuracy gain at 20% of features: **+6.8 %**
  - Biggest accuracy loss at 20% of features: **-20%**
  - Across all experiments, change in accuracy **from -5% to +5%** of validation performance
- **RFE** as a feature selection technique:
  - Highest accuracy gain at 20% of features: **+26 %**
  - Biggest accuracy loss at 20% of features: **-12.3%**
  - Across all experiments, change in accuracy **from -10% to +10%** of validation performance
  - Low feature selection stability between different estimators
- **PCA** and **FA** techniques:
  - Change in accuracy for **PCA: from -22% to +4%** of validation performance
  - Change in accuracy for **FA: from -10% to +6%** of validation performance
  - PCA + adaptive boosting is the best (+ 2% **Diff. Val Acc. b/w PCA and total**)
  - FA + gradient boosting / adaptive boosting is the best (+ 3% **Diff. Val Acc. b/w PCA and total /** (+ 1% **Diff. Val Acc. b/w FA and total)**
- 3 **model-independent, data-independent feature selectors**:
  - Chi2 selector
  - Mutual information selector
  - ANOVA testing selector

Backups