Advanced Machine Learning- Project 2

MICHAŁ GROMADZKI

Preprocessing

- Dataset is split into train and validation with proportions 80:20
- Removing features with high correlation
- StandardScaler from sklearn

Algorithm for removing variables with high correlation:

- Algorithm is iteration-based, in each iteration 1 variable can be removed
- In each iteration:
  - Calculate the correlation between all variable pairs and take the absolute value of it
  - Exclude correlation equal to 1, occurs on the diagonal of the correlation matrix
  - Remove one of the features from the pair with the highest correlation
- Algorithm stops when there are no variable pairs with correlation over TH = 0.9

Custom Metric

Custom metric is calculated as follows:

1. Calculate the probability of belonging to class 1 for all observations in the validation set
2. Select 20% of observations with the highest probability
3. Calculate how many observations were correctly selected - corr
4. Calculate the final value of the metric:

$$corr \cdot 5 \cdot 10 - 200 \cdot num$$

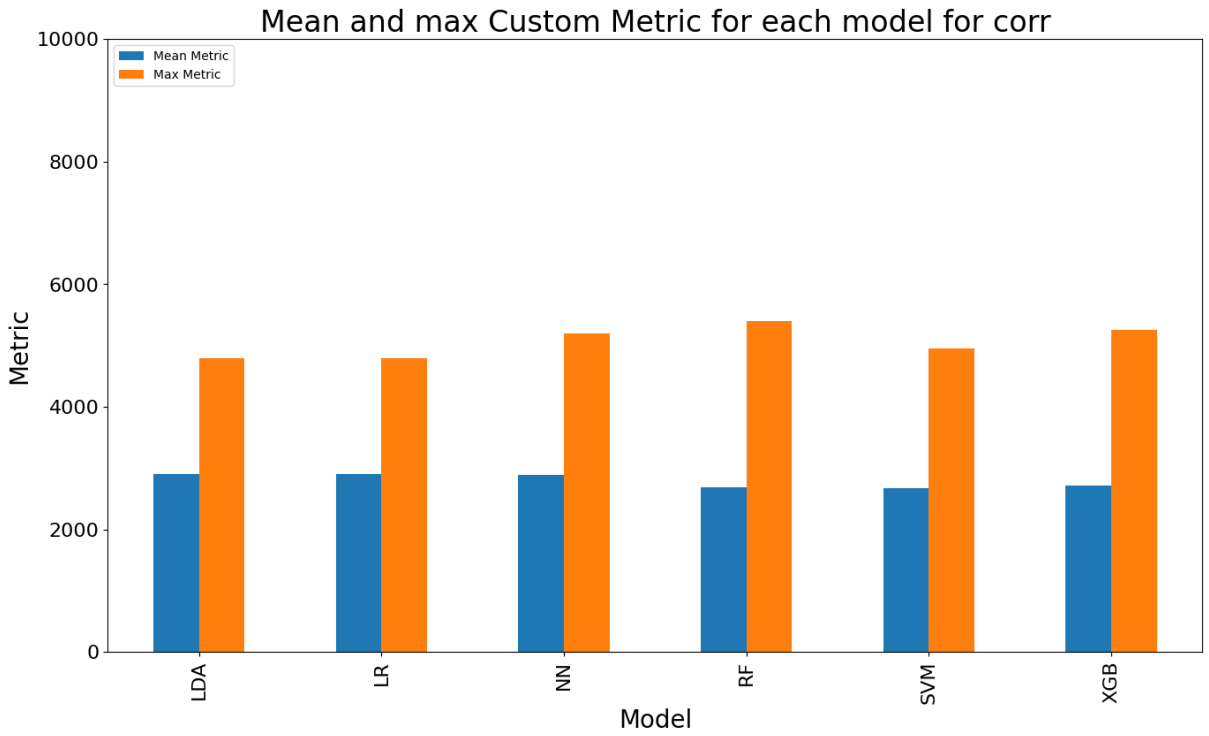
where num is the number of used features

Selected Models

- NN - Neural network
- LR - Logistic Regression
- XGB - XGBoost
- RF - Random Forest
- SVC - Support Vector Machine
- LDA - Linear Discriminant Analysis

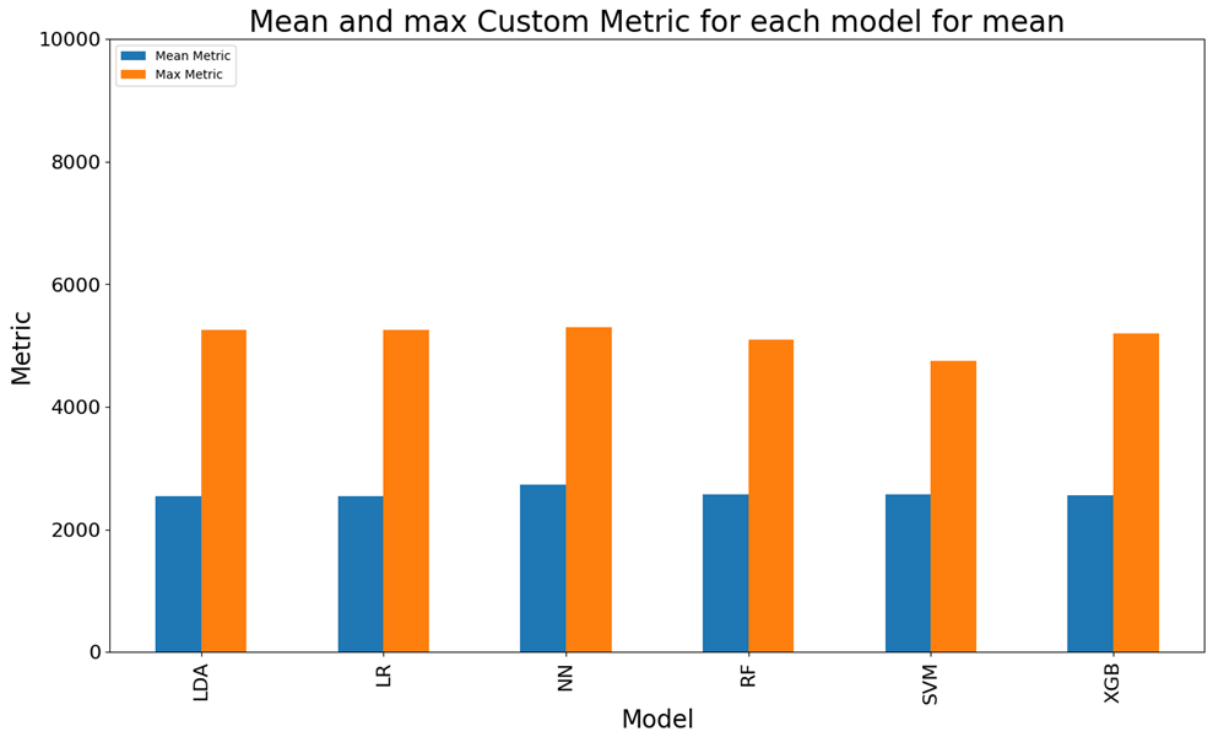
Correlation

Firstly, calculate the correlation of all features to the target variable. Then sort all the variables in descending order based on the calculated correlation. Lastly, in each iteration select n first features from the ordered list



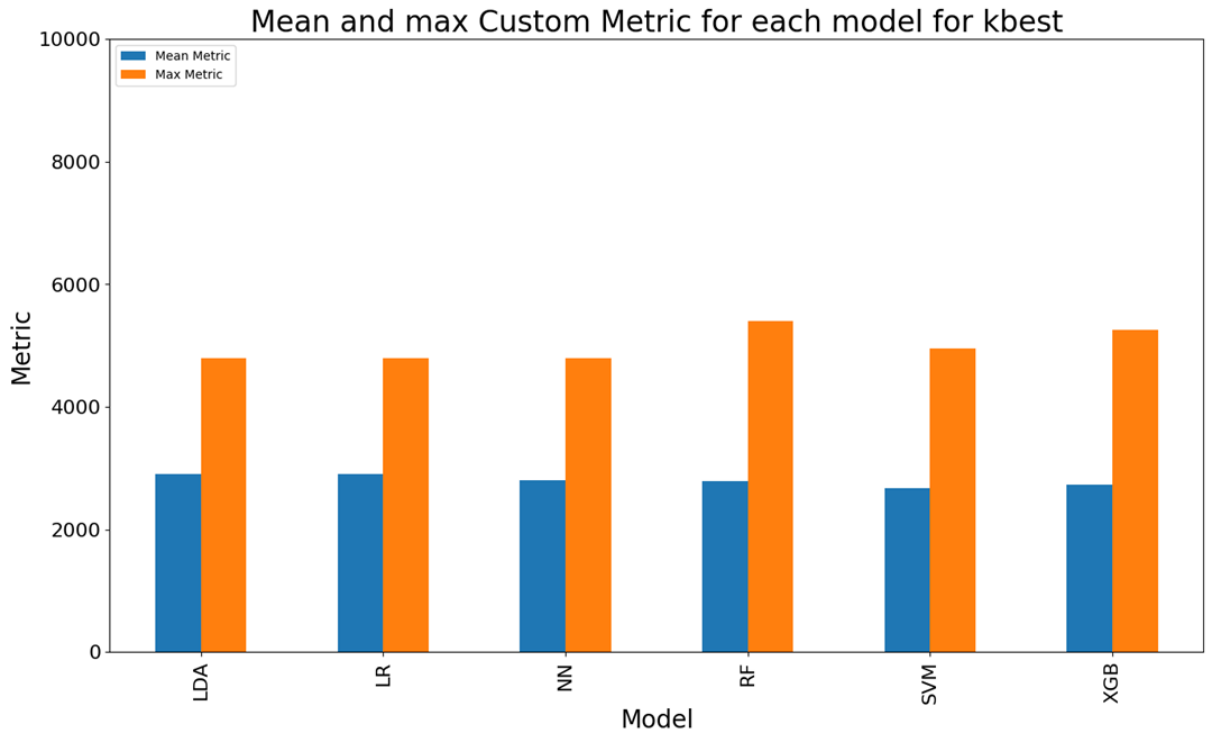
Mean

Firstly, it calculates the means for all features in each target variable class. Then it calculates the absolute value of the difference between both means and divides it by the means of the feature calculated on the entire dataset.



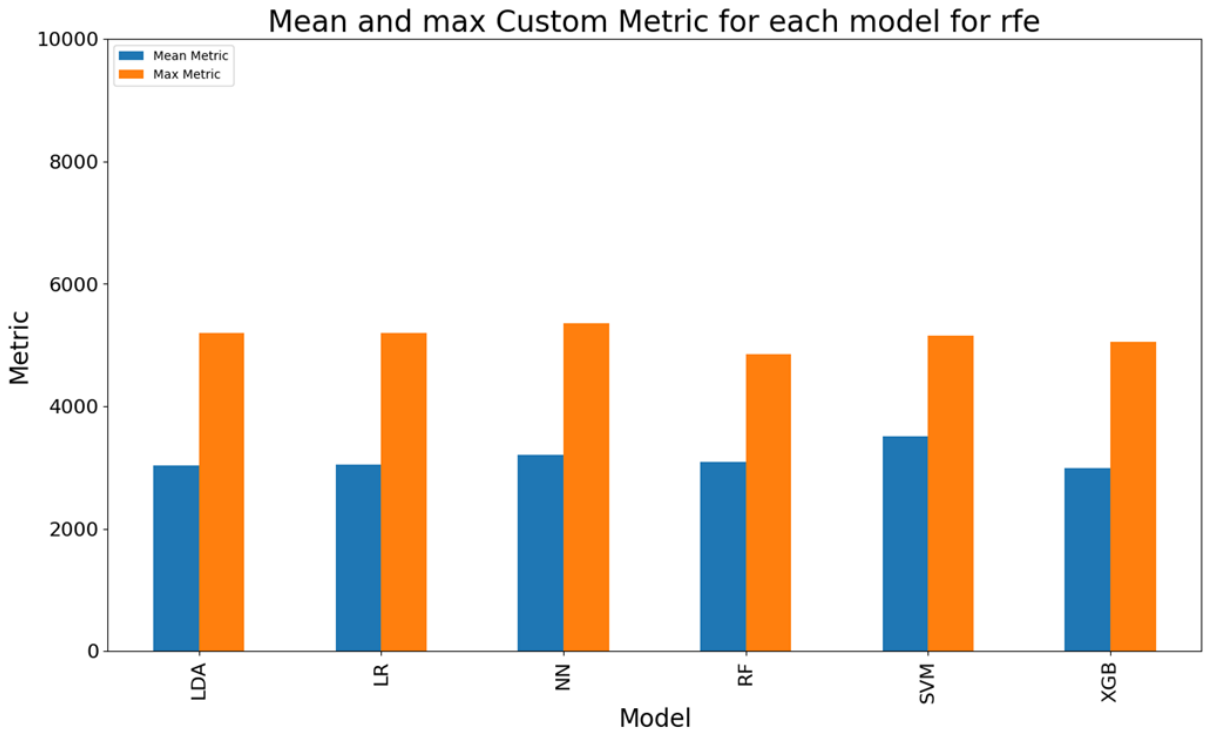
SelectKBest

SelectKBest feature selection method selects k best features according to some scoring function, where the user sets k. In this case, the scoring function is ANOVA F-value.



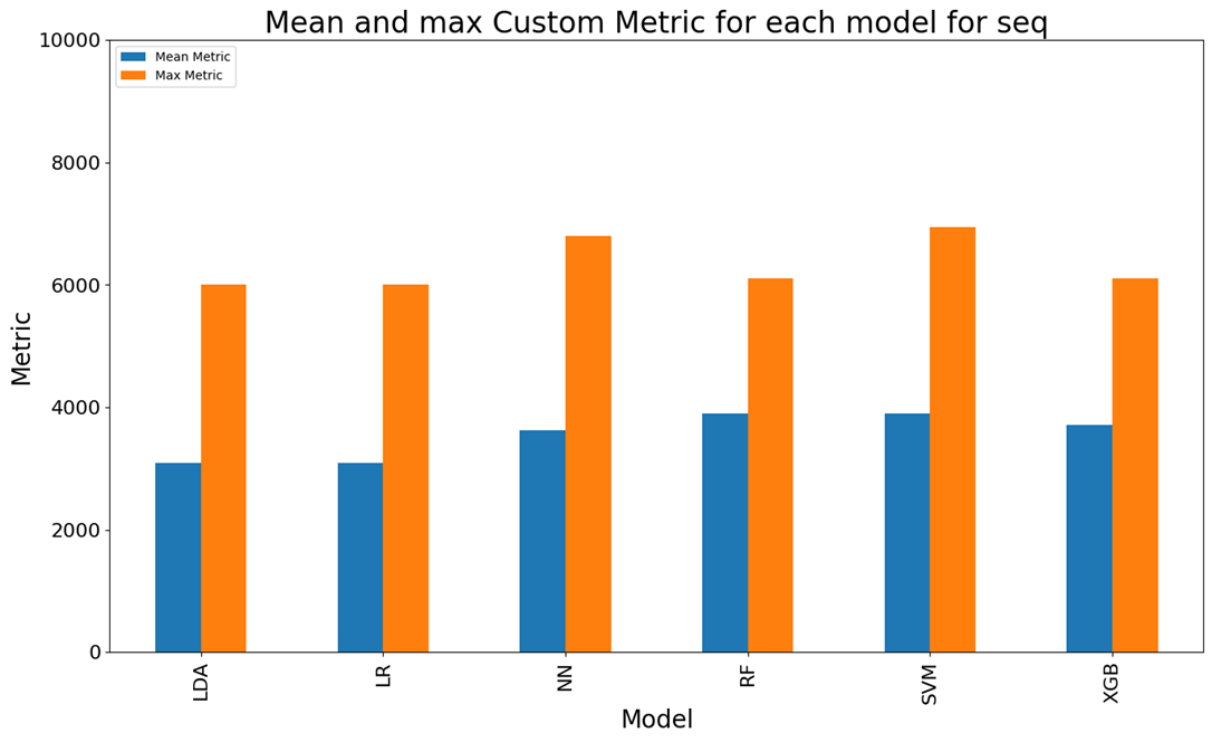
Recursive Feature Elimination

First, the estimator is trained on the initial set of features and the importance of each feature is obtained. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.



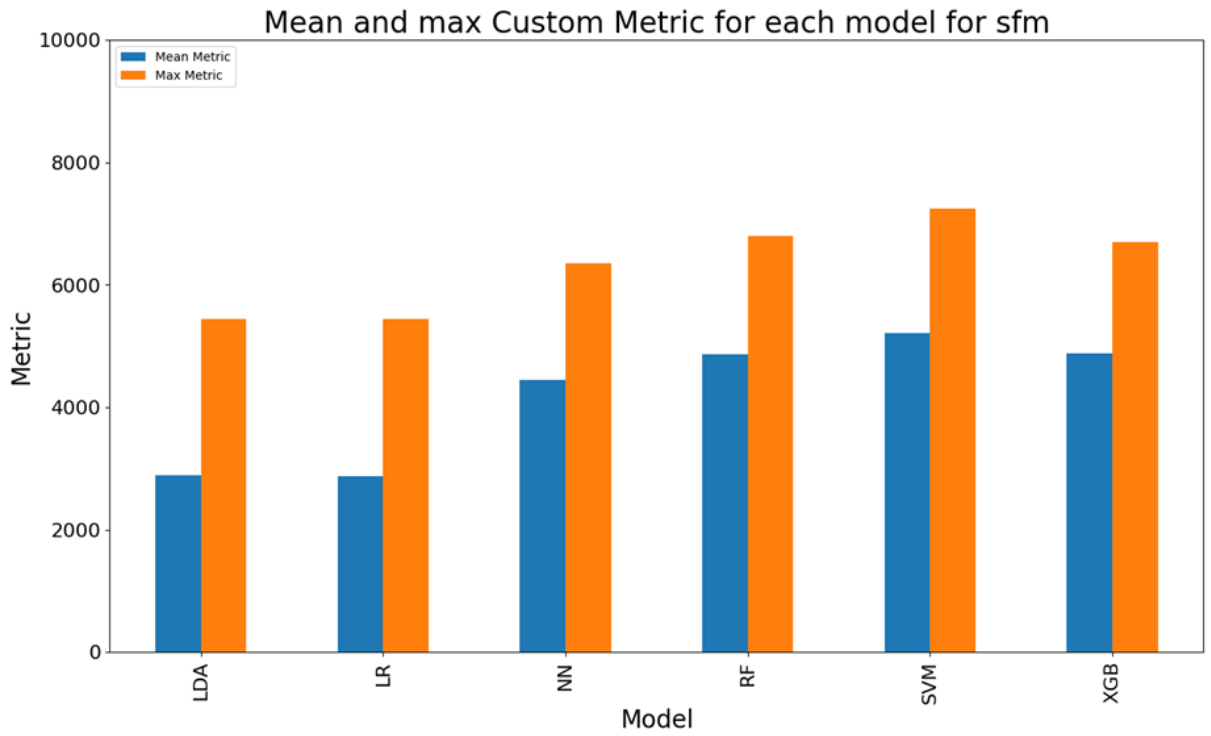
SequentialFeatureSelector

This Sequential Feature Selector adds features to form a feature subset in a greedy fashion. At each stage, this estimator chooses the best feature to add or remove based on the cross-validation score of an estimator.



SelectFromModel

This method selects features based on the feature importance calculated by a provided estimator. In this case, RandomForest was used.



Results – Custom Metric

	NN	LR	XGB	RF	SVM	LDA
Corr	2888	2902	2710	2694	2674	**2904**
Mean	**2728**	2544	2558	2576	2578	2542
KBest	2800	2902	2730	2792	2674	**2904**
SFM	4452	2876	4876	4864	**5214**	2884
Seq	3630	3092	3718	3898	**3906**	3094
RFE	3212	3042	2990	3092	**3516**	3040

Table 3: Mean custom metric for each model for each method

	NN	LR	XGB	RF	SVM	LDA
Corr	5200	4800	5250	**5400**	4950	4800
Mean	**5300**	5250	5200	5100	4750	5250
KBest	4800	4800	5250	**5400**	4950	4800
SFM	6350	5450	6700	6800	**7250**	5450
Seq	6800	6000	6100	6100	**6950**	6000
RFE	**5350**	5200	5050	4850	5150	5200

Table 4: Max custom metric for each model for each method

Results - accuracy

	NN	LR	XGB	RF	SVM	LDA
Corr	0.520	**0.529**	0.507	0.513	0.508	**0.529**
Mean	**0.517**	0.515	0.511	0.510	0.498	0.515
KBest	0.521	**0.529**	0.508	0.514	0.508	**0.529**
SFM	0.623	0.510	0.638	0.657	**0.677**	0.510
Seq	0.554	0.513	0.557	**0.566**	**0.566**	0.513
RFE	0.542	0.536	0.530	0.536	**0.550**	0.536

Table 1: Mean accuracy for each model for each method

	NN	LR	XGB	RF	SVM	LDA
Corr	**0.550**	0.546	0.533	0.533	0.526	0.546
Mean	0.542	0.531	**0.546**	0.532	0.536	0.532
KBest	0.533	**0.546**	0.528	0.543	0.526	**0.546**
SFM	0.685	0.530	0.663	0.694	**0.724**	0.529
Seq	**0.599**	0.541	0.579	0.586	0.598	0.541
RFE	**0.608**	0.556	0.601	0.601	0.598	0.557

Table 2: Max accuracy for each model for each method

Final model selection

- 350 models with 1 to 5 features tested, SelectFromModel method
- Hyper-parameters tuning
- Interactions between features
- SVM

	num_features	columns	accuracy	CustomMetric	C	kernel	degree	Interactions
1st Model	4	100, 102, 103, 105	0.639	**7400**	0.1	poly	4	No
2nd Model	5	100, 102, 103, 104, 105	**0.679**	7300	0.001	linear	3	Yes

Table 5: Metrics and hyper-parameters for two best-performing models

Conclusion

- SVM:
  - C – 0.001
  - Kernel – linear
  - Degree – 3
  - Gamma - auto
- Features – 100, 102, 103, 104, 105; indexed from 0
- With interactions
- Trained on the entire dataset

THE END