Tabular Data Science - Final Project

Eden Yosub, 212784128

March 13, 2023

Abstract

Features selection is a method of choosing a subset of the most relevant and important features out of the original features group. This project goal is to improve this element in the DS pipeline. While feature selection can has many advantages like improving the model explainability, it can also impair the performance of the model by removing necessary information. my intention was to create a new method that combines feature selection methods with feature extraction methods in order to improve performance without harming the model explainability. From the experiment I conducted and the comparison of quality indicators of both the accuracy of the model and its explainability, it appears that for 2 out of 4 datasets, the combined method I created improved the model performance without affecting explainability too much. In 1 dataset there was no affect over performances at all and in the final dataset the model accuracy slightly decreased when I used the combined method. Overall, the proposed method shows signs of improvement compared to the existing methods and can be effective in certain cases.

1 Problem description

In my project I intend to focus on the effects of Feature selection on the model performances and will try to improve this element in the DS pipeline. As we learned in class, there are various methods for feature selection: methods for calculating correlations between different columns and statistical tests like the chi-squared test that can help choose the most relevant features that have strong correlation with the target column. Feature selection can have many benefits: It can help with preventing overfitting, improve runtime, reduce model complexity, improve model explainability and improve the model performance by removing irrelevant and non-informative features.

However, feature selection methods can miss more complex relations between the features to the target. For example, if features X1 and X2 are not correlative to the target feature T ,but have a linear or other connection to the target feature such as: $T = X_1 * B_1 + X_2 * B_2$.

The feature selection method will remove these features and by that can impair performance. one way to overcome this problem is to use different method of dimension reduction called feature extraction. Unlike feature selection where we choose a subset of features from the initial ones, feature extraction combines the original features and creates a set of brand new features.

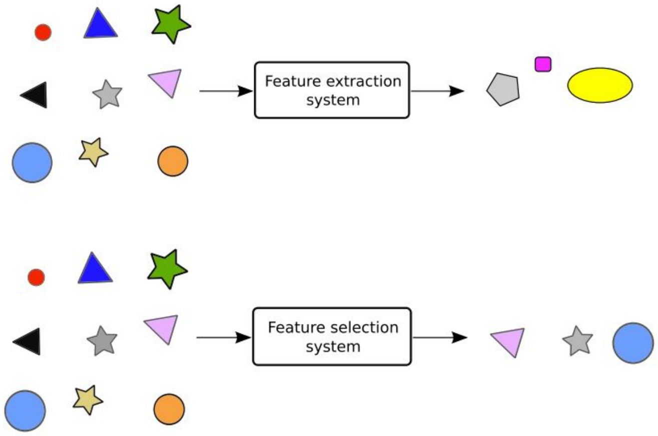

Figure 1: the difference between feature selection and feature extraction

however, feature extraction extract new features that are a combination of the original features. These features may have no physical meaning, therefore we won't be able to explain the model prediction ('model explainability') with them, an ability that can be very important in certain models and tasks.
In my project I intend to explore a way to combine these methods in order to create one method that can has good performance along to high explainabilty.

2 Solution overview

2.1 General approach

The combined method starts by extracting new features using feature extraction methods (PCA and ICA). After I have the new features I will combine them with the original features. Afterwards I will choose the most important features from the original and new ones using feature selection method (correlation + chi squared). I tried different combinations of the methods: extraction algorithm, correlation type (spearman vs pearson), number of features to extract, number of feature to select etc.

2.2 Feature extraction

I tried both ICA and PCA methods, The most practical difference between both techniques is that PCA is useful for finding a reduced-rank representation of the data. ICA, on the other hand, is for finding independent sub-elements in the data. Before combining this new features with the originals, I wanted to check their performance by themselves. I extracted different amount of features using each of the methods. we can see for example a visualization of extracting 3 features with PCA algorithm, the different colors represents the different classes:

The results were not consistent, in some cases PCA worked better and in some ICA worked better. For that reason, in the final method I extracted features using both methods

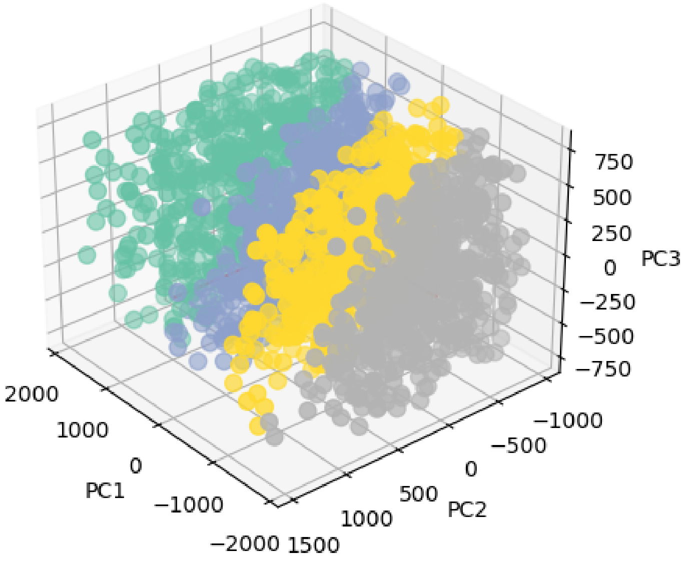

Figure 2: PCA (n = 3),mobile price dataset

and choosing between them using correlation selection.

2.3 Feature selection

I tested both Spearman and Pearson methods for feature selection. The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables as the Spearman Coefficient works with monotonic relationships as well. I chose to use Spearman because unlike Pearson, it is focus on detecting whether as one variable increases, the other variable tends to increase or decrease as well, even if it is not the same increase or decrease in both features. I also tested different numbers of feature to select, in my final method 20% of the features remaining and the other 80% are being ignored.

3 Experimental evaluation

3.1 Data sets

I tried to choose data sets from various domains and types that have a large number of features in order to examine the feature selection methods. Below are the selected datasets:

1. House Prices - the data set that we saw in class. This data contains 81 columns with details about The properties. The task is to predict the house price (regression).

2. Breast Cancer - this data contains 32 Features that was computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The task is to predict if the breast tissues is malignant or benign (binary classification).

3. mushrooms - this data contains 23 columns with various parameters of the mushrooms (color, odor, shape etc). The task is to predict whether the mushroom is poisonous or not (binary classification).

4. Mobile Price - this data contains 21 columns with details about The mobile devices. The task is to predict the mobile price range between 4 classes, low to high cost (multi class classification).

3.2 model performance

the performance evaluation is divided into two parts:

1. for regression datasets, we will usually use The classic method of r^2 score. This method will not fit in this case because as we saw in class, r^2 score of two models with different number of features is incomparable (r^2 increases for adding a feature X_i even if the resulted B_i is insignificant. therefore I won't be able to measure the effect of my feature selection methods. instead, I will use the Adjusted R Squared (AJR) measure which address this issue.

2. for classification dataset I will use the sklearn accuracy score.

For each of the datasets I will use the same evaluation score in order to be able to compare between the different methods.

Data Set	no feature selection	Feature selection (20%)	combined method
House Prices	0.109	0.850	0.847
Breast Cancer	0.528	0.877	0.906
mushrooms	1.000	0.976	0.976
Mobile Price	0.616	0.628	0.750

As we can see, for 2 out of 4 datasets the new method improved the results.

3.3 model explainability

model is explainable when we can point to several features that had the most impact on the model results. For example We would like to be able to say: "this apartment is expensive because it is big, new and received a high rating in the general score feature".
We need to measure this in a numeric score in order to compere the different methods. I wasn't able to find someone that created this kind of measurement system yet, so I defined a new scoring method. The idea of the scoring method is that the model gets a higher score if it can be explained by only a low number of features (which I defined to be 5).
Linear models can use their coefficients as a metric for the overall importance of each feature, but they are scaled with the scale of the variable itself, which might lead to distortions and misinterpretations. Also, the coefficient cannot account for the local importance of the feature,

and how it changes with lower or higher values. therefor I use the SHAP library that we saw in class instead. The score will be the average of the scores for the test set where each score is the the ratio of the five highest SHAP values compared to all the SHAP values.
If this average is high, it means that the first five features give the most information for each prediction, otherwise the prediction is built on the basis of too many features and therefore is not informative enough. extracted features (features that created using extraction algorithm) won't count in the top 5 features. For example if all top 5 features are extracted ones, the explainability score will be 0.

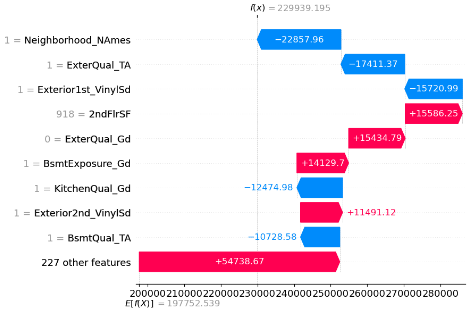

Figure 3: low explainability

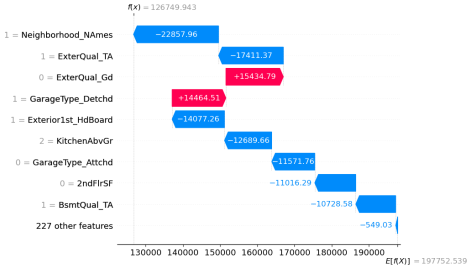

Figure 4: high explainability

Data Set	no feature selection	Feature selection (20%)	combined method
House Prices	0.278	0.459	0.398
Breast Cancer	1.000	1.000	0.993
mushrooms	0.442	0.632	0.631
Mobile Price	0.959	1.000	0.993

We can see that the feature selection method improve the expalinability of the model as expected, while the combined method slightly adversely affects the expalinability compared to the selection method. This is because the new features cannot count as explainable features. Therefor if they are the most important features with out any original feature, the model can't be explained.

4 Related work

There are many methods that are used for performing feature selection. We saw some of them in class, like Correlation based feature selection and Chi-square Statistics based feature selection. There are also some other methods for feature selection that I didn't examine in this project such as using SVM classifier and using unsupervised methods.
When I started to explore the subject of feature selection for this project, I came across an article from the medical field, that compared between feature selection and feature extraction. Their conclusion was that feature extraction is not a good approach in respect of readability, interpretability and transparency - characteristics necessary for the development of trustworthy artificial intelligence, especially in the medical field. they chose to use feature selection

methods instead, even if this comes at the cost of losing some accuracy.
this paper inspired me to try to create a method that adress this problem. In order to do that, I used existing methods like PCA and ICA Feature Extraction.
my method is different by trying to combine these two methods together and by put emphasis on model explainability.

5 Conclusion

This paper presents a new method for feature selection using feature extraction. This approach was able to achieve good results compared to the existing methods without a high cost of loosing to much explainability. I think this can be a window to improve explanatory power without compromising results. Personally, as someone who already experimented a little bit with data science projects, I am so happy about the learning process I experienced in this project. I feel that this project helped me improve my research skills by encouraging me to look at some other work done in the same field and compare my approach to different methods. I also think that the fact that I needed to compare my method over 4 different datasets helped me understand better the processes that the data goes through and the correct way to deal with different cases.