

[Start Here](#)[Blog](#)[Books](#)[About](#)[Contact](#)

Search...



Need help with Python Machine Learning? [Take the FREE Mini-Course](#)

Feature Selection For Machine Learning in Python

by **Jason Brownlee** on May 20, 2016 in **Python Machine Learning**



The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

Irrelevant or partially relevant features can n

In this post you will discover [automatic feature selection](#) to reduce the size of your machine learning data in python with scikit-learn.

Let's get started.

Update Dec/2016: Fixed a typo in the RFE section. Thanks to [David Anderson](#).

Get Your Start in Machine Learning



You can master applied Machine Learning **without the math or fancy degree.**

Find out how in this *free* and *practical* email course.

[START MY EMAIL COURSE](#)



Feature Selection For Machine Learning in Python
Photo by [Baptiste Lafontaine](#), some rights reserved.

Feature Selection

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having irrelevant features in your data can do harm to linear algorithms like linear and logistic regression.

Three benefits of performing feature selection are:

- **Reduces Overfitting:** Less redundant data and less noise.
- **Improves Accuracy:** Less misleading data.
- **Reduces Training Time:** Less data means faster training.

You can learn more about feature selection in the book:

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree.

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

Need help with Machine Learning in Python?

Take my free 2-week email course and discover data prep, algorithms and more (with sample code).

Click to sign-up now and also get a free PDF Ebook version of the course.

Start Your FREE Mini-Course Now!

Feature Selection for Machine Learning

This section lists 4 feature selection recipes for machine learning in Python

This post contains recipes for feature selection methods.

Each recipe was designed to be complete and standalone so that you can copy-and-paste it directly into you project and use it immediately.

Recipes uses the [Pima Indians onset of diabetes dataset](#) to demonstrate the feature selection method. This is a binary classification problem where all of the attributes are numeric.

1. Univariate Selection

Statistical tests can be used to select those features that have the strongest relationship with the output variable.

The scikit-learn library provides the [SelectKBest](#) class to use statistical tests to select a specific number of features.

The example below uses the chi squared (chi2) test to select the top 4 of the best features from the Pima Indians dataset.

```
1 # Feature Extraction with Univariate Selection
2 import pandas
3 import numpy
4 from sklearn.feature_selection import SelectKBest
5 from sklearn.feature_selection import chi2
6 # load data
7 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.csv"
8 names = ['preg', 'plas', 'pres', 'sk', 'ins', 'bmi', 'fam', 'ped', 'age', 'yob', 'diab', 'out']
9 dataframe = pandas.read_csv(url, names=names)
10 array = dataframe.values
```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

```

11 X = array[:,0:8]
12 Y = array[:,8]
13 # feature extraction
14 test = SelectKBest(score_func=chi2, k=4)
15 fit = test.fit(X, Y)
16 # summarize scores
17 numpy.set_printoptions(precision=3)
18 print(fit.scores_)
19 features = fit.transform(X)
20 # summarize selected features
21 print(features[0:5,:])

```

You can see the scores for each attribute and the 4 attributes chosen (those with the highest scores): *plas*, *test*, *mass* and *age*.

```

1 [ 111.52  1411.887   17.605   53.108  2175.565  127.669    5.393
2    181.304]
3 [[ 148.    0.    33.6   50. ]
4 [  85.    0.    26.6   31. ]
5 [ 183.    0.    23.3   32. ]
6 [  89.   94.    28.1   21. ]
7 [ 137.  168.   43.1   33. ]]

```

2. Recursive Feature Elimination

The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain.

It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

You can learn more about the [RFE](#) class in the scikit-learn documentation.

The example below uses RFE with the logistic regression algorithm to select the top 3 features. The choice of algorithm does not matter too

```

1 # Feature Extraction with RFE
2 from pandas import read_csv
3 from sklearn.feature_selection import
4 from sklearn.linear_model import Log
5 # load data
6 url = "https://archive.ics.uci.edu/m
7 names = ['preg', 'plas', 'pres', 'sk
8 dataframe = read_csv(url, names=name
9 array = dataframe.values
10 X = array[:,0:8]
11 Y = array[:,8]
12 # feature extraction
13 model = LogisticRegression()
14 rfe = RFE(model, 3)
15 fit = rfe.fit(X, Y)
16 print("Num Features: %d" % fit.n_fe

```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

```
17 print("Selected Features: %s") % fit.support_
18 print("Feature Ranking: %s") % fit.ranking_
```

You can see that RFE chose the the top 3 features as *preg*, *mass* and *pedi*.

These are marked True in the *support_* array and marked with a choice "1" in the *ranking_* array.

```
1 Num Features: 3
2 Selected Features: [ True False False False False  True  True False]
3 Feature Ranking: [1 2 3 5 6 1 1 4]
```

3. Principal Component Analysis

Principal Component Analysis (or PCA) uses linear algebra to transform the dataset into a compressed form.

Generally this is called a data reduction technique. A property of PCA is that you can choose the number of dimensions or principal component in the transformed result.

In the example below, we use PCA and select 3 principal components.

Learn more about the PCA class in scikit-learn by reviewing the [PCA API](#). Dive deeper into the math behind PCA on the [Principal Component Analysis Wikipedia article](#).

```
1 # Feature Extraction with PCA
2 import numpy
3 from pandas import read_csv
4 from sklearn.decomposition import PCA
5 # load data
6 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes
7 names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
8 dataframe = read_csv(url, names=names)
9 array = dataframe.values
10 X = array[:,0:8]
11 Y = array[:,8]
12 # feature extraction
13 pca = PCA(n_components=3)
14 fit = pca.fit(X)
15 # summarize components
16 print("Explained Variance: %s") % fi
17 print(fit.components_)
```

You can see that the transformed dataset (3 source data.

```
1 Explained Variance: [ 0.88854663  0.0
2 [[ -2.02176587e-03  9.78115765e-02
3    9.93110844e-01  1.40108085e-02
4    [  2.26488861e-02  9.72210040e-01
5    -9.46266913e-02  4.69729766e-02
6    [ -2.24649003e-02  1.43428710e-01
```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

```
7 2.09773019e-02 -1.32444542e-01 -6.39983017e-04 -1.25454310e-01]]
```

4. Feature Importance

Bagged decision trees like Random Forest and Extra Trees can be used to estimate the importance of features.

In the example below we construct a `ExtraTreesClassifier` classifier for the Pima Indians onset of diabetes dataset. You can learn more about the [ExtraTreesClassifier](#) class in the scikit-learn API.

```
1 # Feature Importance with Extra Trees Classifier
2 from pandas import read_csv
3 from sklearn.ensemble import ExtraTreesClassifier
4 # load data
5 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes"
6 names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
7 dataframe = read_csv(url, names=names)
8 array = dataframe.values
9 X = array[:,0:8]
10 Y = array[:,8]
11 # feature extraction
12 model = ExtraTreesClassifier()
13 model.fit(X, Y)
14 print(model.feature_importances_)
```

You can see that we are given an importance score for each attribute where the larger score the more important the attribute. The scores suggest at the importance of *plas*, *age* and *mass*.

```
1 [ 0.11070069  0.2213717  0.08824115  0.08068703  0.07281761  0.14548537  0.12654214  0.
```

Summary

In this post you discovered feature selection for preparing machine learning data in Python with scikit-learn.

You learned about 4 different automatic feature selection methods:

- Univariate Selection.
- Recursive Feature Elimination.
- Principle Component Analysis.
- Feature Importance.

If you are looking for more information on feature selection, you can check out these resources:

- [Feature Selection with the Caret R Package](#)
- [Feature Selection to Improve Accuracy](#)

Get Your Start in Machine Learning

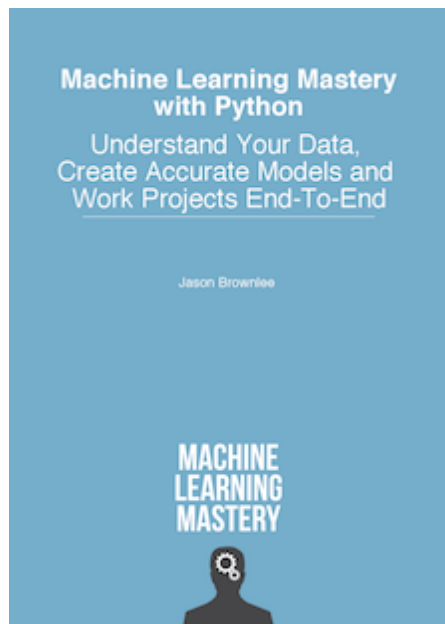
You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

- [An Introduction to Feature Selection](#)
- [Feature Selection in Python with Scikit-Learn](#)

Do you have any questions about feature selection or this post? Ask your questions in the comment and I will do my best to answer them.

Frustrated With Python Machine Learning?



Develop Your Own Models in Minutes

...with just a few lines of scikit-learn code

Discover how in my new Ebook:

[Machine Learning Mastery With Python](#)

Covers **self-study tutorials** and **end-to-end projects** like:
Loading data, visualization, modeling, tuning, and much more...

Finally Bring Machine Learning To Your Own Projects

Skip the Academics. Just Results.

[Click to learn more.](#)



About Jason Brownlee

Dr. Jason Brownlee is a husband, father, developer and a machine learning enthusiast who has started and get good at applied machine learning.

[View all posts by Jason Brownlee](#) →

Get Your Start in Machine Learning




You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

< How To Build Multi-Layer Perceptron Neural Network Models with Keras

Evaluate the Performance of Machine Learning Algorithms in Python using Resampling >


122 Responses to *Feature Selection For Machine Learning in Python*



Juliet September 16, 2016 at 8:57 pm #

REPLY ↩


Hi Jason! Thanks for this – really useful post! I’m sure I’m just missing something simple, but looking at your Univariate Analysis, the features you have listed as being the most correlated seem to have the highest values in the printed score summary. Is that just a quirk of the way this function outputs results? Thanks again for a great access-point into feature selection.



Jason Brownlee September 17, 2016 at 9:29 am #

REPLY ↩


Hi Juliet, it might just be coincidence. If you uncover something different, please let me know.



Ansh October 11, 2016 at 12:16 pm #


REPLY ↩

For the Recursive Feature Elimination, are the features of high importance (preg,mass,pedi)?
The ranking array has value 1 for them then



Jason Brownlee October 12, 2016 at 12:25 pm #

Hi Ansh, I believe the features you mentioned are the first ranked features in the post. These are the first ranked features in the post.



Ansh October 12, 2016 at 12:25 pm #

Thanks for the reply Jason.

Get Your Start in Machine Learning

✕

You can master applied Machine Learning without the math or fancy degree . Find out how in thisfree and practicalemail course.

REPLY ↩

START MY EMAIL COURSE



Jason Brownlee October 13, 2016 at 8:33 am #

REPLY ↩

No problem Ansh.



Anderson Neves December 15, 2016 at 6:52 am #

Hi all,

I agree with Ansh. There are 8 features and the indexes with True and 1 match with preg, mass and pedi.

```
[ 'preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age' ]  
[ True, False, False, False, False, True, True, False]  
[ 1, 2, 3, 5, 6, 1, 1, 4 ]
```

Jason, could you explain better how you see that preg, pedi and age are the first ranked features?

Thank you for the post, it was very useful and direct to the point. Congratulations.



Jason Brownlee December 15, 2016 at 8:31 am #

Hi Anderson, they have a “true” in their column index and are all ranked “1” at their respective column index.

Does that help?



Anderson Neves

Hi Jason,

That is exactly what I mean.
and age in the scenario below.

Features:

```
[ 'preg', 'plas', 'pres', 'skin', 't
```

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

RFE result:

```
[ True, False, False, False, False, False, True, True ]
```

```
[ 1, 2, 3, 5, 6, 4, 1, 1 ]
```

However, the result was

Features:

```
[ 'preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age' ]
```

RFE result:

```
[ True, False, False, False, False, True, True, False]
```

```
[ 1, 2, 3, 5, 6, 1, 1, 4 ]
```

Did you consider the target column 'class' by mistake?

Thank you for the quick reply,

Anderson Neves



Jason Brownlee December 16, 2016 at 5:48 am #

Hi Anderson,

I see, you're saying you have a different result when you run the code?

The code is correct and does not include the class as an input.

Re-running now I see the same result:

```
1 Num Features: 3
2 Selected Features: [ True False False False False True True False]
3 Feature Ranking: [1 2 3 5 6 1 1 4]
```

Perhaps I don't understand /



Anderson Neves

Hi Jason,

Your code is correct and my
features found with RFE are
"You can see that RFE chose
add the code below at the e

```
# find best features
best_features = []
```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

```
i = 0
for is_best_feature in fit.support_:
    if is_best_feature:
        best_features.append(names[i])
    i += 1
print '\nSelected features:'
print best_features
```

Sorry if I am bothering somehow,
Thanks again,
Anderson Neves



Jason Brownlee December 17, 2016 at 11:18 am #

Got it Anderson.
Thanks for being patient with me and helping to make this post more useful. I really appreciate it!
I've fixed up the example above.



Narasimman October 14, 2016 at 9:18 pm #

REPLY ↩

from the rfe, how do I form a new dataframe for the features which has true value?



Jason Brownlee October 15, 2016 at 10:18 am #

Great question Narasimman.

From memory, you can use `numpy.concatenate` to join arrays.
<http://docs.scipy.org/doc/numpy/reference>

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

REPLY ↩



Iain Dinwoodie November 1, 2016 at 10:18 am #

Thanks for useful tutorial.

Narasimman – 'from the rfe, how do I form a new dataframe for the features which has true value?'

START MY EMAIL COURSE

You can just apply rfe directly to the dataframe then select based on columns:

```
...
df = read_csv(url, names=names)
X = df.iloc[:, 0:8]
Y = df.iloc[:, 8]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, 3)
fit = rfe.fit(X, Y)
print("Num Features: {}".format(fit.n_features_))
print("Selected Features: {}".format(fit.support_))
print("Feature Ranking: {}".format(fit.ranking_))

X = X[X.columns[fit.support_]]
```



MLBeginner October 25, 2016 at 1:07 am #

REPLY ↩

Hi Jason,

Really appreciate your post! Really great! I have a quick question for the PCA method. How to get the column header for the selected 3 principal components? It is just simple column no. there, but hard to know which attributes finally are.

Thanks,



Jason Brownlee October 25, 2016 at 8:29 am #

REPLY ↩

Thanks MLBeginner, I'm glad y

There is no column header, they are "ne
helps.



sadiq October 25, 2016 at 1:51 am #

hi, Jason! please I want to ask you
analysis by python

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

Jason Brownlee October 25, 2016 at 8:29 am #

REPLY ↩



Sure, try it and see how the results compare (as in the models trained on selected features) to other feature selection methods.



Vignesh Sureshababu Kishore November 15, 2016 at 5:07 pm #

REPLY ↩

Hey Jason, can the univariate test of Chi2 feature selection be applied to both continuous and categorical data.



Jason Brownlee November 16, 2016 at 9:25 am #

REPLY ↩

Hi Vignesh, I believe just continuous data. But I may be wrong – try and see.



Vignesh Sureshababu Kishore November 16, 2016 at 1:07 pm #

REPLY ↩

Hey Jason, Thanks for the reply. In the univariate selection to perform the chi-square test you are fetching the array from `df.values`. In that case, each element of the array will be each row in the data frame.

To perform feature selection, we should have ideally fetched the values from each column of the dataframe to check the independence of each feature with the class variable. Is it a inbuilt functionality of the `sklearn.preprocessing` because of which you fetch the values as each row.

Please suggest me on this.



Jason Brownlee Nov

I'm not sure I follow Vignesh to perform the tests.



Vineet December 2, 2016 at 5:11 am #

REPLY ↩

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE



Hi Jason,

I am trying to do image classification using cpu machine, I have very large training matrix of 3800*200000 means 200000 features. Pls suggest how do I reduce my dimension.?



Jason Brownlee December 2, 2016 at 8:19 am #

REPLY ↩

Consider working with a sample of the dataset.

Consider using the feature selection methods in this post.

Consider projection methods like PCA, sammons mapping, etc.

I hope that helps as a start.



tvmanikandan December 15, 2016 at 5:49 pm #

REPLY ↩

Jason,

when you use "SelectKBest" , can you please explain how you get the below scores?

[111.52 1411.887 17.605 53.108 2175.565 127.669 5.393
181.304]

-Mani



Jason Brownlee December 16, 2016 at 5:40 am #

REPLY ↩

I use a chi squared test, you ca

http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE



tvmanikandan December 16, 2016 at 5:29 pm #

REPLY ↩

Jason,

I understand you used chi square. But if want to get these scores manually , how can i do it? Please explain.

-Mani



Jason Brownlee December 17, 2016 at 11:05 am #

REPLY ↩

Good question, I don't have an example at the moment sorry.



tvmanikandan December 16, 2016 at 2:48 am #

REPLY ↩

jason,

Please explain how the below scores are achieved using chi2.

[111.52 1411.887 17.605 53.108 2175.565 127.669 5.393
181.304]

-Mani



Natheer Alabsi December 28, 2016 at 8:35 pm #

REPLY ↩

Jason, how can we get feature nam



Jason Brownlee December 29, 2016 at 11:05 am #

Hi Natheer,

Map the feature rank to the index of the
or whathaveyou.



Jason January 9, 2017 at 2:40 am #

REPLY ↩

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

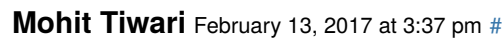
START MY EMAIL COURSE



I have a regression problem and I need to convert a bunch of categorical variables into dummy data, which will generate over 200 new columns. Should I do the feature selection before this step or after this step?

REPLY

That is a lot of new binary variables. Your resulting dataset will be sparse (lots of zeros). Feature selection prior might be a good idea, also try after.



REPLY ↩

Can u please suggest me a suitable feature



START MY EMAIL COURSE

REPLY



Hi Mohit,

Consider trying a few different methods, as well as some projection methods and see which “views” of your data result in more accurate predictive models.



Esu February 15, 2017 at 12:01 am #

REPLY ↩

Hell!

Once I got the reduced version of my data as a result of using PCA, how can I feed to my classifier?

example: the original data is of size 100 row by 5000 columns
if I reduce 200 features I will get 100 by 200 dimension data. right?
then I create arrays of

```
a=array[:,0:199]
```

```
b=array[:,99]
```

but when I test my classifier its core is 0% in both test and training accuracy?

An7y Idea



Jason Brownlee February 15, 2017 at 11:35 am #

REPLY ↩

Sounds like you're on the right, but a zero accuracy is a red flag.

Did you accidentally include the class output variable in the data when doing the PCA? It should be excluded.



Kamal February 20, 2017 at 6:20 pm #

Hello sir,
I have a question in my mind
each of these feature selection algo uses so
we come to know that my data set contain o
automatically select no features its own.

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

REPLY ↩



Jason Brownlee February 21, 2017 at 11:35 am #



Great question Kamal.

No, you must select the number of features. I would recommend using a sensitivity analysis and try a number of different features and see which results in the best performing model.



Massimo March 9, 2017 at 5:29 am #

REPLY ↩

Hi Jason,

I have a question about the RFECV approach.

I'm dealing with a project where I have to use different estimators (regression models). is it correct use RFECV with these models? or is it enough to use only one of them? Once I have selected the best features, could I use them for each regression model?

To better explain:

- I have used RFECV on whole dataset in combination with one of the following regression models [LinearRegression, Ridge, Lasso]
- Then I have compared the r^2 and I have chosen the better model, so I have used its features selected in order to do others things.
- pratically, I use the same 'best' features in each regression model.

Sorry for my bad english.



Jason Brownlee March 9, 2017 at 9:58 am #

REPLY ↩

Good question.

You can embed different models in RFE and see if the results tell the same or different stories in terms of what features to pick.

You can build a model from each set of features.

You can pick one set of features and build a model.

My advice is to try everything you can think of on the validation dataset.

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.



Massimo March 11, 2017 at 2

Thank you man. You're gre

START MY EMAIL COURSE



Jason Brownlee March 11, 2017 at 8:01 am #

REPLY ↩

You're welcome.



gevra March 22, 2017 at 1:49 am #

REPLY ↩

Hi Jason.

Thanks for the post, but I think going with Random Forests straight away will not work if you have correlated features.

Check this paper:

<https://academic.oup.com/bioinformatics/article/27/14/1986/194387/Classification-with-correlated-features>

I am not sure about the other methods, but feature correlation is an issue that needs to be addressed before assessing feature importance.



Jason Brownlee March 22, 2017 at 8:08 am #

REPLY ↩

Makes sense, thanks for the note and the reference.



ssh June 20, 2017 at 8:20 pm #

REPLY ↩

Jason, following this notes, among the input vectors become pre the features reduction technics which optimization with gradient descent) s Thanks



Jason Brownlee June

Perhaps a correlation a values, select features and use

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

ogunleye March 30, 2017 at 4:29 am #

REPLY ↩

Hello sir,

Thank you for the informative post. My questions are

1) How do you handle NaN in a dataset for feature selection purposes.

2) I am getting an error with RFE(model, 3) It is telling me i supplied 2 arguments instead of 1.

Thank you very much once again.



Jason Brownlee March 30, 2017 at 8:57 am #

REPLY ↩

Hi, NaN is a mark of missing data.

Here are some ways to handle missing data:

<http://machinelearningmastery.com/handle-missing-data-python/>



ogunleye March 30, 2017 at 4:33 am #

REPLY ↩

I solved my problem sir. I named the function RFE in my main but. I would love to hear your response to first question.



Sam April 20, 2017 at 3:49 am #

REPLY ↩

how to load the nested JSON into t



Jason Brownlee April 20, 2017 at 3:49 am #

I don't know off hand, perhaps



Federico Carmona April 20, 2017 at 3:49 am #

good afternoon

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

REPLY ↩

START MY EMAIL COURSE

How to know with pca what are the main components?



Jason Brownlee April 20, 2017 at 9:34 am #

REPLY ↩

PCA will calculate and return the principal components.



Federico Carmona April 20, 2017 at 10:53 am #

REPLY ↩

Yes but pca does not tell me which are the most relevant vars if mass test etc?



Jason Brownlee April 21, 2017 at 8:27 am #

REPLY ↩

Not sure I follow you sorry.

You could apply a feature selection or feature importance method to the PCA results if you wanted. It might be overkill though.



Lehyu April 23, 2017 at 6:44 pm #

REPLY ↩

In RFE we should input a estimator, so before I do feature selection, should I fine tune the model or just use the default parmater setting? Thanks.



Jason Brownlee April 24, 2017 at 10:00 am #

You can, but that is not really recommended. On the problem, the selected features will be the most relevant.



Lehyu April 25, 2017 at 12:41 am #

I was suck here for days. Thanks for your help.

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE



Lehyu April 25, 2017 at 1:09 am #

REPLY ↩

stuck...



Jason Brownlee April 25, 2017 at 7:49 am #

REPLY ↩

I'm glad to hear the advice helped.

I'm here to help if you get stuck again, just post your questions.



Rj May 7, 2017 at 4:38 pm #

REPLY ↩

Hi Jason,

I was wondering if I could build/train another model (say SVM with RBF kernel) using the features from SVM-RFE (wherein the kernel used is a linear kernel).



Jason Brownlee May 8, 2017 at 7:42 am #

REPLY ↩

Sure.



Gwen June 5, 2017 at 7:02 pm #

REPLY ↩

Hi Jason,

First of all thank you for all your posts ! It's v

I'm working on a personal project of prediction accuracy of 65% (not awesome but it's a goal that will affect my predictions. So I applied two algorithms – Recursive Feature Elimination, – Feature Importance.

But I have some contradictions. For example, I found a feature the most important in Feature Importance but it's not the most important in Recursive Feature Elimination ?

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

In addition to that in Feature Importance all features are between 0,03 and 0,06... Is that mean that all features are not correlated with my output ?

Thanks again for your help !



Jason Brownlee June 6, 2017 at 9:30 am #

REPLY ↩

Hi Gwen,

Different feature selection methods will select different features. This is to be expected.

Build a model on each set of features and compare the performance of each.

Consider ensembling the models together to see if performance can be lifted.

A great area to consider to get more features is to use a rating system and use rating as a highly predictive input variable (e.g. chess rating systems can be used directly).

Let me know how you go.



Gwen June 7, 2017 at 1:17 am #

REPLY ↩

Thanks for your answer Jason.

I tried with 20 features selected by Recursive Feature Elimination but my accuracy is about 60%...

In addition to that the Elo Rating system (used in chess) is one of my features. With this feature only my accuracy is ~65%.

Maybe a MLP is not a good idea for
I only have one hidden layer.

And maybe we cannot have more than
(Not enough for a positive ROI !)



Jason Brownlee June

Hang in there Gwen.

Try lots of models and lots of combinations

See what skill other people get out of

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

REPLY ↩

START MY EMAIL COURSE

is possible.

Brainstorm for days about features and other data you could use.

See this post:

<http://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/>



RATNA NITIN PATIL July 20, 2017 at 8:16 pm #

REPLY ↩

Hello Jason,

I am very much impressed by this tutorial. I am just a beginner. I have a very basic question. Once I got the reduced version of my data as a result of using PCA, how can I feed to my classifier? I mean to say how to feed the output of PCA to build the classifier?



Jason Brownlee July 21, 2017 at 9:33 am #

REPLY ↩

Assign it to a variable or save it to file then use the data like a normal input dataset.



RATNA NITIN PATIL July 20, 2017 at 8:56 pm #

REPLY ↩

Hi Jason,

I was trying to execute the PCA but, I got the error at this point of the code

```
print("Explained Variance: %s") % fit.explain
```

It's a type error: unsupported operand type(s)

Please help me.



Jason Brownlee July 21, 2017 at 9:33 am #

Looks like a Python 3 issue. Maybe

```
1 print("Explained Variance: %s")
```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE



RATNA NITIN PATIL July 21, 2017 at 2:23 pm #

REPLY ↩

Thanks Jason. It works.



Jason Brownlee July 22, 2017 at 8:29 am #

REPLY ↩

Glad to hear it.



Raphael Alencar July 21, 2017 at 9:57 pm #

REPLY ↩

How to know wich feature selection technique i have to choose?



Jason Brownlee July 22, 2017 at 8:35 am #

REPLY ↩

Consider using a few, create models for each and select the one that results in the best performing model.



RATNA NITIN PATIL July 22, 2017 at 4:23 pm #

REPLY ↩

Hello Jason,

I have used the extra tree classifier for the fe each attribute. But then I want to provide the the classifier. I am not able to provide only th I would be grateful to you if you help me in th



Jason Brownlee July 23, 2017 at 10:10 am #

The importance scores are for to use as inputs to your model.

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

RATNA NITIN PATIL July 22, 2017 at 6:33 pm #

REPLY ↩

Hi Jason,

Basically i want to provide feature reduction output to Naive Bays. I f you could provide sample code will be better.

Thanks for providing this wonderful tutorial.



Jason Brownlee July 23, 2017 at 6:21 am #

REPLY ↩

You can use feature selection or feature importance to “suggest” which features to use, then develop a model with those features.



RATNA NITIN PATIL July 23, 2017 at 6:44 pm #

REPLY ↩

Thanks Jason,

But after knowing the important features, I am not able to build a model from them. I don't know how to giveonly those feautreslimportant) as input to the model. I mean to say X_train parameter will have all the features as input.

Thanks in advance....



Jason Brownlee July 24, 2017 at 6:21 am #

A feature selection method will suggest which features to use, then develop a model with those features. You can use your favorite programming language to make a model with those features.



RATNA NITIN PATIL July 24, 2017 at 6:21 am #

thanks a lot Jason. You are



Jason Brownlee July 24, 2017 at 6:21 am #

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

REPLY ↩



Thanks.



RATNA NITIN PATIL July 24, 2017 at 6:11 pm #

REPLY ↩

I have my own dataset on the Desktop, not the standard dataset that all machine learning have in their depositories (e.g. iris , diabetes).

I have a simple csv file and I want to load it so that I can use scikit-learn properly.

I need a very simple and easy way to do so.

Waiting for the reply.



Jason Brownlee July 25, 2017 at 9:37 am #

REPLY ↩

Try this tutorial:

<http://machinelearningmastery.com/load-machine-learning-data-python/>



mlearn July 29, 2017 at 6:04 am #

REPLY ↩

Thanks for this post, it's very helpful,

What would make me choose one technique and not the others?

The results of each of these techniques correlates with the result of others?, I mean, makes sense to use more than one to verify the feature selection?.

thanks!



Jason Brownlee July 29, 2017 at 6:04 am #

Choose a technique based on the results of the others.

In predictive modeling we are concerned with decreasing model complexity.

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

REPLY ↩



Sounds that I'd need to cross-validate each technique... interesting, I know that heavily depends on the data but I'm trying to figure out an heuristic to choose the right one, thanks!.



Jason Brownlee July 31, 2017 at 8:14 am #

REPLY ↩

Applied machine learning is empirical. You cannot pick the "best" methods analytically.



steve August 17, 2017 at 3:15 pm #

REPLY ↩

Hi Jason,

In your examples, you write:

```
array = dataframe.values
```

```
X = array[:,0:8]
```

```
Y = array[:,8]
```

In my dataset, there are 45 features. When i write like this:

```
X = array[:,0:44]
```

```
Y = array[:,44]
```

I get some errors:

```
Y = array[:,44]
```

IndexError: index 45 is out of bounds for axis 1 with size 0

If you help me, i ll be grateful!

Thanks in advance.



Jason Brownlee August 17, 2017 at 3:15 pm #

Confirm that you have loaded y



Aneeshaa S C August 20, 2017 at 11:20 am #

1.. What kind of predictors can be u

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

REPLY ↩

START MY EMAIL COURSE

2. If categorical predictors can be used, should they be re-coded to have numerical values? ex: yes/no values re-coded to be 1/0
3. Can categorical variables such as location (U(urban)/R(rural)) be used without any conversion/re-coding?



Jason Brownlee August 21, 2017 at 6:07 am #

REPLY ↩

Regression, e.g. predicting a real value.

Categorical inputs must be encoded as integers or one hot encoded (dummy variables).



panteha August 29, 2017 at 1:36 am #

REPLY ↩

Hi Jason

I am new to ML and am doing a project in Python, at some point it is to recognize correlated features , I wonder what will be the next step? what to do with correlated features? should we change them to something new? a combination maybe? how does it affect our modeling and prediction? appreciated if you direct me into some resources to study and find it out.
best



Jason Brownlee August 29, 2017 at 5:09 pm #

REPLY ↩

It is common to identify and remove the correlated input variables.

Try it and see if it lifts skill on your model



Silvio Abela September 26, 2017 at 6:4

Hello Dr Brownlee,

Thank you for these incredible tutorials.

I am trying to classify some text data collected from a dataset. Is there any way in which the constants in the model can be removed? For example, in SelectKBest, k=3, in RFE you can specify the number of features. Feature Importance it is left open for selection that way

Get Your Start in Machine Learning



REPLY ↩

You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

Is there a way like a rule of thumb or an algorithm to automatically decide the “best of the best”? Say, I use n-grams; if I use trigrams on a 1000 instance data set, the number of features explodes. How can I set SelectKBest to an “x” number automatically according to the best? Thank you.



Jason Brownlee September 26, 2017 at 2:59 pm #

REPLY ↩

No, hyperparameters cannot be set analytically. You must use experimentation to discover the best configuration for your specific problem.

You can use heuristics or copy values, but really the best approach is experimentation with a robust test harness.



Abby October 6, 2017 at 3:43 pm #

REPLY ↩

It was an impressive tutorial, quite easy to understand. I am looking for feature subset selection using gaussian mixture clustering model in python. Can you help me out?



Jason Brownlee October 7, 2017 at 5:48 am #

REPLY ↩

Sorry, I don't have material on mixture models or clustering. I cannot help.



Manjunat October 6, 2017 at 8:31 pm #

REPLY ↩

Hi jason

I've tried all feature selection techniques while modelling ...?



Jason Brownlee October 7, 2017 at 10:00 am #

Try many for your dataset and select the best model.

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

Nerea October 16, 2017 at 7:13 pm #

REPLY ↩

Hello Jason,

I am a biochemistry student in Spain and I am on a project about predictive biomarkers in cancer. The bioinformatic method I am using is very simple but we are trying to predict metastasis with some protein data. In our research, we want to determine the best biomarker and the worst, but also the synergic effect that would have the use of two biomarkers. That is my problem: I don't know how to calculate which are the two best predictors.

This is what I have done for the best and worst predictors:

```
analysis=['il10meta']
X = data[analysis].values

#response variable
response='evol'
y = data[response].values

# use train/test split with different random_state values
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=5)

from sklearn.neighbors import KNeighborsClassifier

#creating the classifier
knn = KNeighborsClassifier(n_neighbors=1)

#fitting the classifier
knn.fit(X_train, y_train)

#predicting response variables corresponding to test data
y_pred = knn.predict(X_test)

#calculating classification accuracy
print(metrics.accuracy_score(y_test, y_pred))
```

I have calculate the accuracy. But when I try result in all the combinations of my 6 biomar

Could you help me? Any tip?

THANK YOU



Jason Brownlee October 17, 20

Generally, I would recommend predictive modeling problem:

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

REPLY ↩

START MY EMAIL COURSE

<https://machinelearningmastery.com/start-here/#process>

Generally, you must test many different models and many different framings of the problem to see what works best.



gen October 17, 2017 at 6:35 pm #

REPLY ↩

Hello Jason,

Many thanks for your post. I have also read your introduction article about feature selection. Which method is Feature Importance categorized under? i.e wrapper or embedded ?

Thanks



Jason Brownlee October 18, 2017 at 5:32 am #

REPLY ↩

Neither, it is a different thing yet again.

You could use the importance scores as a filter.



Numan Yilmaz October 26, 2017 at 1:46 pm #

REPLY ↩

Great post! Thank you, Jason. My question is all these in the post here are integers. That is needed for all algorithms. What if I have categorical data? How can I know which feature is more important for the model if there are categorical features? Is there a method/way to calculate it before one-hot encoding(get_dummies) or is not tree-based?



Jason Brownlee October 26, 2017 at 1:46 pm #

Good question, I cannot think of a method off hand, they may be out there. See my search).



rohit November 13, 2017 at 9:11 pm #

REPLY ↩

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE



hello Jason,

Should I do Feature Selection on my validation dataset also? Or just do feature selection on my training set alone and then do the validation using the validation set?



Jason Brownlee November 14, 2017 at 10:10 am #

REPLY ↩

Use the train dataset to choose features. Then, only choose those features on test/validation and any other dataset used by the model.



Maryam November 16, 2017 at 3:45 pm #

REPLY ↩

hello jason

i am doing simple classification but there is coming an issue

ValueError Traceback (most recent call last)

in ()

—> 1 fit = test.fit(X, Y)

~\Anaconda3\lib\site-packages\sklearn\feature_selection\univariate_selection.py in fit(self, X, y)

339 Returns self.

340 """

-> 341 X, y = check_X_y(X, y, ['csr', 'csc'], multi_output=True)

342

343 if not callable(self.score_func):

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in check_X_y(X, y, accept_sparse, dtype, order, copy, force_all_finite, ensure_2d, allow_nd, multi_output, ensure_min_samples,

ensure_min_features, y_numeric, warn_on_

571 X = check_array(X, accept_sparse, dtype,

572 ensure_2d, allow_nd, ensure_min_sam

-> 573 ensure_min_features, warn_on_dtype

574 if multi_output:

575 y = check_array(y, 'csr', force_all_finite=

~\Anaconda3\lib\site-packages\sklearn\utils\

dtype, order, copy, force_all_finite, ensure_2

ensure_min_features, warn_on_dtype, estim

431 force_all_finite)

432 else:

-> 433 array = np.array(array, dtype=dtype,

434

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

435 if ensure_2d:

ValueError: could not convert string to float: 'no'
can you guide me in this regard



Jason Brownlee November 17, 2017 at 9:19 am #

REPLY ↩

You may want to use a label encoder and a one hot encoder to convert string data to numbers.



Vinod P November 17, 2017 at 12:19 am #

REPLY ↩

```
import numpy as np
from pandas import read_csv
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
# load data
data = read_csv('C:\\Users\\abc\\Downloads\\xyz\\api.csv', names =
['org.apache.http.impl.client.DefaultHttpClient.execute','org.apache.http.impl.client.DefaultHttpClie
nt.','java.net.URLConnection.getInputStream','java.net.URLConnection.connect','java.net.URL.ope
nStream','java.net.URL.openConnection','java.net.URL.getContent','java.net.Socket.','java.net.Serv
erSocket.bind','java.net.ServerSocket.','java.net.HttpURLConnection.connect','java.net.DatagramS
ocket.','android.widget.VideoView.stopPlayback','android.widget.VideoView.start','android.widget.Vi
deoView.setVideoURI','android.widget.VideoView.setVideoPath','android.widget.VideoView.pause',
'android.text.format.DateUtils.formatDateTime','android.text.format.DateFormat.getTimeFormat','a
ndroid.text.format.DateFormat.getDateForm
d.telephony.TelephonyManager.getSubscrib
alNumber','android.telephony.TelephonyMar
anager.getLine1Number','android.telephony.
extToSpeech.','android.provider.Settings$Sy
nt','android.provider.Settings$System.getCol
g','android.provider.Settings$Secure.getInt','
','android.os.PowerManager$WakeLock.rele
ndroid.net.wifi.WifiManager.setWifiEnabled',
et.wifi.WifiManager.getWifiState','android.net
fiManager.getConnectionInfo','android.media
one.play','android.media.MediaRecorder.set
d.media.MediaPlayer.start','android.media.M
er.reset','android.media.MediaPlayer.release
```

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .
Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

```
MediaPlayer.pause','android.media.MediaPlayer.create','android.media.AudioRecord.','android.location.LocationManager.requestLocationUpdates','android.location.LocationManager.removeUpdates','android.location.LocationManager.getProviders','android.location.LocationManager.getLastKnownLocation','android.location.LocationManager.getBestProvider','android.hardware.Camera.open','android.bluetooth.BluetoothAdapter.getAddress','android.bluetooth.BluetoothAdapter.enable','android.bluetooth.BluetoothAdapter.disable','android.app.WallpaperManager.setBitmap','android.app.KeyguardManager$KeyguardLock.reenableKeyguard','android.app.KeyguardManager$KeyguardLock.disableKeyguard','android.app.ActivityManager.killBackgroundProcesses','android.app.ActivityManager.getRunningTasks','android.app.ActivityManager.getRecentTasks','android.accounts.AccountManager.getAccountsByType','android.accounts.AccountManager.getAccounts','Class']])
```

```
dataframe = read_csv(url, names=names)
array = dataframe.values
X = array[:,0:70]
Y = array[:,70]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, 3)
fit = rfe.fit(X, Y)
#print("Num Features: %d") % fit.n_features_
#print("Selected Features: %s") % fit.support_
#print("Feature Ranking: %s") % fit.ranking_
```

I get following error

ValueError Traceback (most recent call last)

in ()

6 model = LogisticRegression()

7 rfe = RFE(model, 3)

—> 8 fit = rfe.fit(X, Y)

9 print("Num Features: %d") % fit.n_features_

10 print("Selected Features: %s") % fit.support_



Jason Brownlee November 17, 2017 at 12:29 am

Perhaps try posting your code to the forum



Vinod P November 17, 2017 at 12:29 am

Get Your Start in Machine Learning

You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

REPLY ↩



Can you post a code on first select relevant features using any feature selection method, and then use relevant features to construct classification model?



Jason Brownlee November 17, 2017 at 9:26 am #

REPLY ↩

Thanks for the suggestion.



Hemalatha December 1, 2017 at 2:12 am #

REPLY ↩

will you post a code on selecting relevant features using feature selection method and then using relevant features constructing a classification model??



Jason Brownlee December 1, 2017 at 7:40 am #

REPLY ↩

Yes, see this post:

<https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/>



Arjun December 13, 2017 at 4:45 am #

REPLY ↩

Hi Jason,
Thanks for the content, it was really helpful.
Can you clarify if the above mentioned meth



Jason Brownlee December 13, 2017 at 4:45 am #

Perhaps, I'm no sure off hand.



Danilo December 25, 2017 at 1:12 pm #

Hi Jason,

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

REPLY ↩

START MY EMAIL COURSE

I just had the same question as Arjun, I tried with a regression problem but neither of the approaches were able to do it.



Jason Brownlee December 25, 2017 at 5:25 am #

REPLY ↩

What was the problem exactly?

Leave a Reply

Name (required)

Email (will not be published) (required)

Website

SUBMIT COMMENT

Welcome to Machine Learning Mastery



Hi, I'm Dr. Jason Brownlee.
My goal is to make practitione

[Read More](#)

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

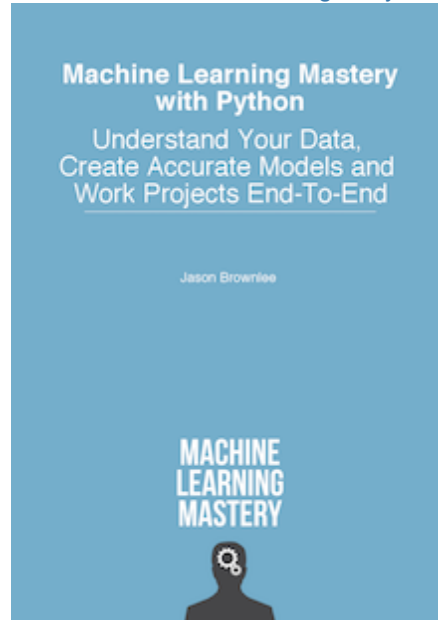
Develop Predictive Models With Python

Want to develop your own models in scikit-learn?

Want step-by-step tutorials?

Looking for sample code and templates?

[Get Started With Machine Learning in Python Today!](#)



POPULAR



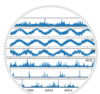
Your First Machine Learning Project in Python Step-By-Step

JUNE 10, 2016



Time Series Prediction with LSTM Recu

JULY 21, 2016



Multivariate Time Series Forecasting wi

AUGUST 14, 2017



Develop Your First Neural Network in Py

MAY 24, 2016



How to Setup a Python Environment for

MARCH 13, 2017

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree .

Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE

**Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras**

JULY 26, 2016

**Time Series Forecasting with the Long Short-Term Memory Network in Python**

APRIL 7, 2017

**Multi-Class Classification Tutorial with the Keras Deep Learning Library**

JUNE 2, 2016

**Regression Tutorial with the Keras Deep Learning Library in Python**

JUNE 9, 2016

**How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras**

AUGUST 9, 2016

© 2018 Machine Learning Mastery. All Rights Reserved.

[Privacy](#) | [Contact](#) | [About](#)

Get Your Start in Machine Learning



You can master applied Machine Learning without the math or fancy degree . Find out how in this *free* and *practical* email course.

START MY EMAIL COURSE