Want help with machine learning? Take the FREE Crash-Course.

10 Standard Datasets for Practicing Applied Machine Learning

by **Jason Brownlee** on November 25, 2016 in **Machine Learning Process**

The key to getting good at applied machine learning is practicing on lots of different datasets.

This is because each problem is different, requiring subtly different data preparation and modeling methods.

In this post, you will discover 10 top standard machine learning datasets that you can use for practice.

Let's dive in.

Overview

A structured Approach

Each dataset is summarized in a consistent way. This makes them easy to compare and navigate for you to practice a specific data preparation technique or modeling method.

The aspects that you need to know about each dataset are:

1. **Name**: How to refer to the dataset.
2. **Problem Type**: Whether the problem is regression or classification.
3. **Inputs and Outputs**: The numbers and known names of input and output features.

4. **Performance**: Baseline performance for comparison using the Zero Rule algorithm, as well as best known performance (if known).
5. **Sample**: A snapshot of the first 5 rows of raw data.
6. **Links**: Where you can download the dataset and learn more.

Standard Datasets

Below is a list of the 10 datasets we'll cover.

Each dataset is small enough to fit into memory and review in a spreadsheet. All datasets are comprised of tabular data and no (explicitly) missing values.

1. Swedish Auto Insurance Dataset.
2. Wine Quality Dataset.
3. Pima Indians Diabetes Dataset.
4. Sonar Dataset.
5. Banknote Dataset.
6. Iris Flowers Dataset.
7. Abalone Dataset.
8. Ionosphere Dataset.
9. Wheat Seeds Dataset.
10. Boston House Price Dataset.

1. Swedish Auto Insurance Dataset

The Swedish Auto Insurance Dataset involves predicting the total payment for all claims in thousands of Swedish Kronor, given the total number of claims.

It is a regression problem. It is comprised of 63 observations with 1 input variable and one output variable. The variable names are as follows:

1. Number of claims.
2. Total payment for all claims in thousands of Swedish Kronor.

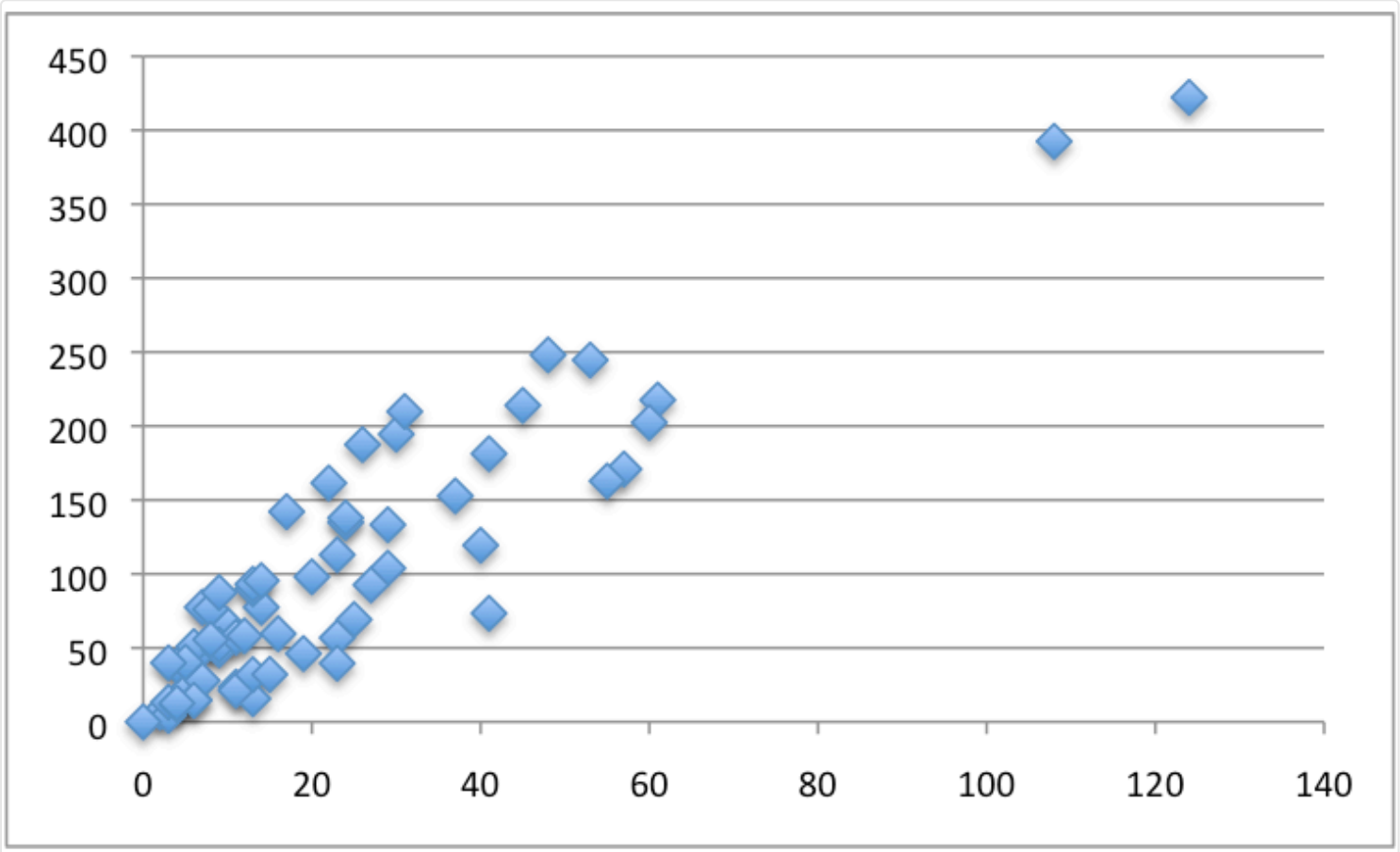
The baseline performance of predicting the mean value is an RMSE of approximately 72.251 thousand Kronor.

A sample of the first 5 rows is listed below.

```
108,392.5
19,46.2
```

```
3 13,15.7
4 124,422.2
5 40,119.4
```

Below is a scatter plot of the entire dataset.



Swedish Auto Insurance Dataset

- Download
- More Information

2. Wine Quality Dataset

The Wine Quality Dataset involves predicting the quality of white wines on a scale given chemical measures of each wine.

It is a multi-class classification problem, but could also be framed as a regression problem. The number of observations for each class is not balanced. There are 4,898 observations with 11 input variables and one output variable. The variable names are as follows:

1. Fixed acidity.

2. Volatile acidity.
3. Citric acid.
4. Residual sugar.
5. Chlorides.
6. Free sulfur dioxide.
7. Total sulfur dioxide.
8. Density.
9. pH.
10. Sulphates.
11. Alcohol.
12. Quality (score between 0 and 10).

The baseline performance of predicting the mean value is an RMSE of approximately 0.148 quality points.

A sample of the first 5 rows is listed below.

```
7,0.27,0.36,20.7,0.045,45,170,1.001,3,0.45,8.8,6
6.3,0.3,0.34,1.6,0.049,14,132,0.994,3.3,0.49,9.5,6
8.1,0.28,0.4,6.9,0.05,30,97,0.9951,3.26,0.44,10.1,6
7.2,0.23,0.32,8.5,0.058,47,186,0.9956,3.19,0.4,9.9,6
7.2,0.23,0.32,8.5,0.058,47,186,0.9956,3.19,0.4,9.9,6
```

- Download
- More Information

3. Pima Indians Diabetes Dataset

The Pima Indians Diabetes Dataset involves predicting the onset of diabetes within 5 years in Pima Indians given medical details.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 768 observations with 8 input variables and 1 output variable. Missing values are believed to be encoded with zero values. The variable names are as follows:

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.

8. Age (years).
9. Class variable (0 or 1).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 65%. Top results achieve a classification accuracy of approximately 77%.

A sample of the first 5 rows is listed below.

```
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
```

- Download
- More Information
- Top Results

4. Sonar Dataset

The Sonar Dataset involves the prediction of whether or not an object is a mine or a rock given the strength of sonar returns at different angles.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 208 observations with 60 input variables and 1 output variable. The variable names are as follows:

1. Sonar returns at different angles
2. …
3. Class (M for mine and R for rock)

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 53%. Top results achieve a classification accuracy of approximately 88%.

A sample of the first 5 rows is listed below.

```
0.0200,0.0371,0.0428,0.0207,0.0954,0.0986,0.1539,0.1601,0.3109,0.2111,0.1609,0.1582,0.2238,0.
0.0453,0.0523,0.0843,0.0689,0.1183,0.2583,0.2156,0.3481,0.3337,0.2872,0.4918,0.6552,0.6919,0.
0.0262,0.0582,0.1099,0.1083,0.0974,0.2280,0.2431,0.3771,0.5598,0.6194,0.6333,0.7060,0.5544,0.
0.0100,0.0171,0.0623,0.0205,0.0205,0.0368,0.1098,0.1276,0.0598,0.1264,0.0881,0.1992,0.0184,0.
0.0762,0.0666,0.0481,0.0394,0.0590,0.0649,0.1209,0.2467,0.3564,0.4459,0.4152,0.3952,0.4256,0.
```

- Download
- More Information

- Top Results

5. Banknote Dataset

The Banknote Dataset involves predicting whether a given banknote is authentic given a number of measures taken from a photograph.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 1,372 observations with 4 input variables and 1 output variable. The variable names are as follows:

1. Variance of Wavelet Transformed image (continuous).
2. Skewness of Wavelet Transformed image (continuous).
3. Kurtosis of Wavelet Transformed image (continuous).
4. Entropy of image (continuous).
5. Class (0 for authentic, 1 for inauthentic).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 50%.

A sample of the first 5 rows is listed below.

```
3.6216,8.6661,-2.8073,-0.44699,0
4.5459,8.1674,-2.4586,-1.4621,0
3.866,-2.6383,1.9242,0.10645,0
3.4566,9.5228,-4.0112,-3.5944,0
0.32924,-4.4552,4.5718,-0.9888,0
4.3684,9.6718,-3.9606,-3.1625,0
```

- Download
- More Information

6. Iris Flowers Dataset

The Iris Flowers Dataset involves predicting the flower species given measurements of iris flowers.

It is a multi-class classification problem. The number of observations for each class is balanced. There are 150 observations with 4 input variables and 1 output variable. The variable names are as follows:

1. Sepal length in cm.
2. Sepal width in cm.
3. Petal length in cm.

4. Petal width in cm.
5. Class (Iris Setosa, Iris Versicolour, Iris Virginica).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 26%.

A sample of the first 5 rows is listed below.

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
```

- Download
- More Information

7. Abalone Dataset

The Abalone Dataset involves predicting the age of abalone given objective measures of individuals.

It is a multi-class classification problem, but can also be framed as a regression. The number of observations for each class is not balanced. There are 4,177 observations with 8 input variables and 1 output variable. The variable names are as follows:

1. Sex (M, F, I).
2. Length.
3. Diameter.
4. Height.
5. Whole weight.
6. Shucked weight.
7. Viscera weight.
8. Shell weight.
9. Rings.

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 16%. The baseline performance of predicting the mean value is an RMSE of approximately 3.2 rings.

A sample of the first 5 rows is listed below.

```
M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15
M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7
```

```
3 F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9
4 M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10
5 I,0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7
```

- Download
- More Information

8. Ionosphere Dataset

The Ionosphere Dataset requires the prediction of structure in the atmosphere given radar returns targeting free electrons in the ionosphere.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 351 observations with 34 input variables and 1 output variable. The variable names are as follows:

1. 17 pairs of radar return data.
2. …
3. Class (g for good and b for bad).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 64%. Top results achieve a classification accuracy of approximately 94%.

A sample of the first 5 rows is listed below.

```
1 1,0,0.99539,-0.05889,0.85243,0.02306,0.83398,-0.37708,1,0.03760,0.85243,-0.17755,0.59755,-0.4
2 1,0,1,-0.18829,0.93035,-0.36156,-0.10868,-0.93597,1,-0.04549,0.50874,-0.67743,0.34432,-0.6970
3 1,0,1,-0.03365,1,0.00485,1,-0.12062,0.88965,0.01198,0.73082,0.05346,0.85443,0.00827,0.54591,0
4 1,0,1,-0.45161,1,1,0.71216,-1,0,0,0,0,0,0,-1,0.14516,0.54094,-0.39330,-1,-0.54467,-0.69975,1,
5 1,0,1,-0.02401,0.94140,0.06531,0.92106,-0.23255,0.77152,-0.16399,0.52798,-0.20275,0.56409,-0.
```

- Download
- More Information
- Top Results

9. Wheat Seeds Dataset

The Wheat Seeds Dataset involves the prediction of species given measurements of seeds from different varieties of wheat.

It is a binary (2-class) classification problem. The number of observations for each class is balanced. There are 210 observations with 7 input variables and 1 output variable. The variable names are as follows:

1. Area.
2. Perimeter.
3. Compactness
4. Length of kernel.
5. Width of kernel.
6. Asymmetry coefficient.
7. Length of kernel groove

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 28%.

A sample of the first 5 rows is listed below.

```
15.26,14.84,0.871,5.763,3.312,2.221,5.22,1
14.88,14.57,0.8811,5.554,3.333,1.018,4.956,1
14.29,14.09,0.905,5.291,3.337,2.699,4.825,1
13.84,13.94,0.8955,5.324,3.379,2.259,4.805,1
16.14,14.99,0.9034,5.658,3.562,1.355,5.175,1
```

- Download
- More Information

10. Boston House Price Dataset

The Boston House Price Dataset involves the prediction of a house price in thousands of dollars given details of the house and its neighborhood.

It is a regression problem. The number of observations for each class is balanced. There are 506 observations with 13 input variables and 1 output variable. The variable names are as follows:

1. CRIM: per capita crime rate by town.
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of nonretail business acres per town.
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. NOX: nitric oxides concentration (parts per 10 million).
6. RM: average number of rooms per dwelling.
7. AGE: proportion of owner-occupied units built prior to 1940.
8. DIS: weighted distances to five Boston employment centers.
9. RAD: index of accessibility to radial highways.
10. TAX: full-value property-tax rate per $10,000.

11. PTRATIO: pupil-teacher ratio by town.
12. B: 1000(Bk – 0.63)^2 where Bk is the proportion of blacks by town.
13. LSTAT: % lower status of the population.
14. MEDV: Median value of owner-occupied homes in $1000s.

The baseline performance of predicting the mean value is an RMSE of approximately 9.21 thousand dollars.

A sample of the first 5 rows is listed below.

```
0.00632 18.00 2.310 0 0.5380 6.5750 65.20 4.0900 1 296.0 15.30 396.90 4.98 24.00
0.02731 0.00 7.070 0 0.4690 6.4210 78.90 4.9671 2 242.0 17.80 396.90 9.14 21.60
0.02729 0.00 7.070 0 0.4690 7.1850 61.10 4.9671 2 242.0 17.80 392.83 4.03 34.70
0.03237 0.00 2.180 0 0.4580 6.9980 45.80 6.0622 3 222.0 18.70 394.63 2.94 33.40
0.06905 0.00 2.180 0 0.4580 7.1470 54.20 6.0622 3 222.0 18.70 396.90 5.33 36.20
```

- Download
- More Information

Summary

In this post, you discovered 10 top standard datasets that you can use to practice applied machine learning.

Here is your next step:

1. Pick one dataset.
2. Grab your favorite tool (like Weka, scikit-learn or R)
3. See how much you can beat the standard scores.
4. Report your results in the comments below.

About Jason Brownlee

Jason is the editor-in-chief at MachineLearningMastery.com. He is a husband, proud father, academic researcher, author, professional developer and a machine learning practitioner. He has a Masters and PhD in Artificial Intelligence, has published books on Machine Learning and has written operational code that is running in production. Learn more.

View all posts by Jason Brownlee →

‹ Machine Learning Performance Improvement Cheat Sheet

5 Top Machine Learning Podcasts ›

No comments yet.

Leave a Reply

Name (required)

Email (will not be published) (required)

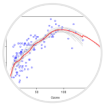
Website

SUBMIT COMMENT

POPULAR

Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras
JULY 21, 2016

How to Run Your First Classifier in Weka
FEBRUARY 17, 2014



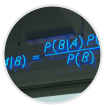
A Tour of Machine Learning Algorithms
NOVEMBER 25, 2013

Develop Your First Neural Network in Python With Keras Step-By-Step
MAY 24, 2016

Tutorial To Implement k-Nearest Neighbors in Python From Scratch
SEPTEMBER 12, 2014

Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras
JULY 26, 2016

Machine Learning for Programmers
AUGUST 17, 2015



How To Implement Naive Bayes From Scratch in Python
DECEMBER 8, 2014

Your First Machine Learning Project in Python Step-By-Step
JUNE 10, 2016

8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset
AUGUST 19, 2015

Privacy | Contact | About