Errol Stoute
Demari Green
Adrian Holley
9/5/22

===========

Baltimore Building Datasets with Solid

What is Data Science?

Data science combines statistics and math, advanced analytics, specialized programming, artificial intelligence, and machine learning to uncover insights hidden in organization's data. Data science involves various tools, roles, and processes which help us extract certain data that we need to do certain tasks. One process being data analysis, which is what we would be using here for our problem. Data analysis helps examine biases, patterns, ranges, and distributions of values within certain data. It also helps with hypothesis generation which leads to results within predictive analytics and machine learning.

What is Machine Learning?

Machine Learning is part of Data Science and is a field of Artificial Intelligence that gives machines human-like capabilities to learn and adapt through statistical models and algorithms. Machine Learning can utilize past datasets in order to perform complex tasks in a way that is similar to how humans solve problems.

Audience Skill Level: All Levels

Our Objective:

The main objective of our project was to make a program that uses information about a building that accurately returns the value of our target. The target feature within our dataset we chose to pick was AREA_ so this would be the main feature or column of values we would try and predict.. To do that we would need to load the data in and read the csv and split the data in test sets so that the machine can give accurate or close to correct predictions.

Errol Stoute
Demari Green
Adrian Holley
9/5/22

What Is Solid?

Solid technology is a specification for allowing users to store their data and keep it in a decentralized data store called **Pods** for private access and keeping it secured. When data is stored in the user's Pod, they control which people and applications can access it.

Solid (derived from "social linked data") is a proposed set of conventions and tools for building decentralized social applications based on Linked Data principles. Solid is modular and extensible and it relies as much as possible on existing W3C standards and protocols. Any info can be stored in solid pods and to store or access data in pods, there's applications that use open, standard, and interoperable data formats and protocols. In terms of the pods, you can get a pod from a pod provider or choose to self-host your own pod and from that, you can have multiple of those pods, even if they aren't the same host. Solid supports storing Linked Data which leads to structuring data as Linked Data means that different applications can work with the same data. Some solid architecture includes HTTP, which they use to transmit data securely over the web and also through components within the pods. Another one is LDP which is a linked data platform which means that pods can contain both private and public containers and each of those with different ACL rules from the user to determine what is shared or public and what is private. In that case, ACLs are files that keep track of who has access to what type of data and when they got access. It's key so just in case they want to revoke access to someone, they can lean towards ACL to figure that out easily.

Errol Stoute
Demari Green
Adrian Holley
9/5/22

NEW PROJECT TOPIC

Nov 02 Meeting - And Work Due

Your task for this week -
https://bniajfi.org/vital_signs/data_downloads/
1) Get a data set about buildings
https://data.baltimorecity.gov/datasets/baltimore::real-property-information-2/explore?location=39.284845%2C-76.620485%2C12.58&showTable=true
https://data.baltimorecity.gov/datasets/baltimore::buildings-footprint/explore?location=39.321833%2C-76.597250%2C10.99&showTable=true - THIS ONE (DATASET)

2) Learn about ETL (Evaluate, Transform, Load).
Definition:
ETL describes the end-to-end process by which a company takes its full breadth of data—structured and unstructured and managed by any number of teams from anywhere in the world—and gets it to a state where it's actually useful for business purposes.
Data cleansing (also known as data scrubbing) is the name of a process of correcting and - if necessary - eliminating inaccurate records from a particular database. The purpose of data cleansing is to detect so called dirty data (incorrect, irrelevant or incomplete parts of the data) to either modify or delete it to ensure that a given set of data is accurate and consistent with other sets in the system.
Source(s): https://www.etltools.org/data-cleansing.html, https://cloud.google.com/learn/what-is-etl

Learn about how to clean data. How to deal with missing values? How to deal with values that are obviously wrong? How to spot and deal with outliers?

And how to do it in Python. **Learn about Python's Tensorflow, dataframes, and the Python Machine Learning API.**

The following is a good place to start --
https://www.w3schools.com/python/python_ml_mean_median_mode.asp

In Particular, you want to know about how to modify your data set to make it more ready for machine learning. This involves data cleaning. Learn how to clean data with Python API.

Here is a good tutorial on dataframes, removing missing values, cleaning data, …
10 minutes to pandas — pandas 1.5.1 documentation (pydata.org)

Errol Stoute
Demari Green
Adrian Holley
9/5/22

3) **For Nov 09** - Concrete Project - ML with Python on building data
Data Cleaning (20% of you project grade)

1. Take your data file.
2. Use the DataFrame API (pandas API) to visualize your data in **Jupyter Notebook** at Welcome To Colaboratory - Colaboratory (google.com)
   a) Show measures of central tendency (You may want to consider converting categorical/nominal values into numerical values) For example, convert low, medium, high values into 1, 2, 3, respectively. Possible reference: Plot With Pandas: Python Data Visualization for Beginners – Real Python
   b) Show diagrams, histograms (nominal values)
   c) Show box plots (whiskers, 25th percentile, 50th percentile, 75th percentile) for each numerical feature or for selected features
   d) Identify any outliers. You can use the boxplot for this.
   e) Plot the correlation between each pair of numeral features
   f) Plot as much as you can of the data. Chart visualization — pandas 1.5.1 documentation (pydata.org)
3. Use the steps at Pythonic Data Cleaning With Pandas and NumPy – Real Python to clean it
   a) May have to remove records with missing values or record that you think are inconsistent
4. Deliverable: Show your JupyterNotebook with the Python code and with the plots and output of the Python commands.

4) Learn about whether or how you can load it on a Solid pod (You'll learn what a Solid pod is)
https://inrupt.com/solid/

I'm waiting for Inrupt to tell us when they'll train us. Look out for a notification from me about Inrupt training, which will go over how to develop an application that interacts with a Solid pod.Such an application is an important component of you main deliverable for your senior project.
--> Important. They will teach us how to write code to manipulate data in our pods.

4) Deliverable:

- The data set(s)
- A python code that loads the data set into a dataframe, and then do ETL on the data that's in the data frame.

Errol Stoute
Demari Green
Adrian Holley
9/5/22

Due before **Nov 16 About (40% of your project done if you do the following also)**

Most Important - Develop Python Code to do machine learning on your data set.

Step 1- Pick one feature of your choosing (i.e., column name to be your target, or label)

Step 2 - I will send you example machine learning implementations where you take your 100,000+ dataset and then do training/testing and tune your model to make accurate predictions on the target. Use these examples to develop and test your own classifier. In other words, you would end with a program *i.e., machine learning model) that would take as input information about a building that would accurately (or to your satisfaction return the value of the target! I'm sure you'd have to keep tuning this program to improve the predictions.

So, your model should give an accurate estimate of that values if you make a brand new reading that's just missing that value.

For Nov 30 - Definitely

Use the dropna() method to drop all columns that are not numbers.

Then you can do

```
# Scatter Plot Matrix
import matplotlib.pyplot as plt
import pandas
from pandas.plotting import scatter_matrix
url =
"https://data.baltimorecity.gov/datasets/baltimore::buildings-footprint.csv"
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = pandas.read_csv(url, names=names)
# Here you do your dropna()
scatter_matrix(data)
plt.show()
```

You can start with this file

JupyterBookForLoadingAndMLonBuildingData.ipynb - Colaboratory (google.com)

Follow the steps at Python Machine Learning Mini-Course - MachineLearningMastery.com

Errol Stoute
Demari Green
Adrian Holley
9/5/22

REFERENCES PAGE

Python Machine Learning Mini-Course - MachineLearningMastery.com

https://bniajfi.org/vital_signs/data_downloads/

Pythonic Data Cleaning With Pandas and NumPy – Real Python

https://catalog.data.gov/dataset/?q=buildings&sort=score+desc%2C+name+asc&groups=local&_organization_limit=0&organization=city-of-baltimore

Chart visualization — pandas 1.5.1 documentation (pydata.org)

Plot With Pandas: Python Data Visualization for Beginners – Real Python

https://data.baltimorecity.gov/datasets/baltimore::real-property-information-2/explore?location=39.284845%2C-76.620485%2C12.58&showTable=true

https://data.baltimorecity.gov/datasets/baltimore::buildings-footprint/explore?location=39.321833%2C-76.597250%2C10.99&showTable=true

https://www.etltools.org/data-cleansing.html
https://towardsdatascience.com/weather-forecasting-with-machine-learning-using-python-55e90c346647

https://www.wunderground.com/history/daily/us/md/baltimore

https://www.tutorialspoint.com/extract-csv-file-specific-columns-to-list-in-python

https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f

https://medium.com/codex/house-price-prediction-with-machine-learning-in-python-cf9df744f7ff