



Python-Powered Machine Learning

Carrie Gardner, PTOH '23

Matthew Fetterman, PTOH '22

Special Thanks:

Jeff Wheelhouse

Data Analytics Club



A Little Background on Us

Carrie Gardner

- Team Lead – Insider Risk @ Software Engineering Institute, CMU
- Social Science & Information Science Education
- 5+ years using python and working on analytics projects
- Proud Python Evangelist 😊

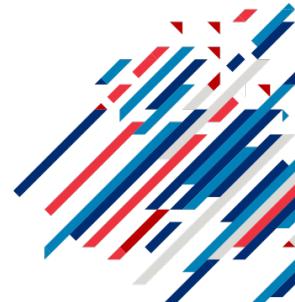
Matthew Fetterman

- Army Officer
- Agriculture & Plant/Soil Science Education
- 10+ years managing analytics projects

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder





Workshop Roadmap

March 17, 2021

- Logistics
- What is Machine Learning?
- Why Python?
- Business Applications
- Interactive Tutorial

Timing:

~30 Minutes Lecture

~5-10 Minute Break

~80 Minutes

Interactive Coding

Logistics

March 17, 2021

- This workshop assumes
 - Attendees have little experience with machine learning
 - Attendees have little experience using python
 - **Everyone** can learn to code (no CS background required!)
- **PLEASE raise your hand and ask questions!**
- Takeaways:
 - **Confidence** to start machine learning with python
 - **Knowledge** of ML and data analytics capabilities with python
 - **Scaffolding** to jump-start a python-powered analytics project





What is Machine Learning?



Let's Frame the Opportunity

March 17, 2021

Dataset

Seattle Airbnb Open Data

A sneak peek into the Airbnb activity in Seattle, WA, USA

Airbnb • updated 3 years ago (Version 2)

Data Tasks (1) Code (60) Discussion (3) Activity Metadata

Download (19 MB) New Notebook

Usability 7.1

License CC0: Public Domain

Tags travel, hotels and accommodations, united states

- Can we **forecast** and **predict** prices?
- Can we **parse** and **glean insights** from reviews?
- Can we **segment** listings into groups?

kaggle

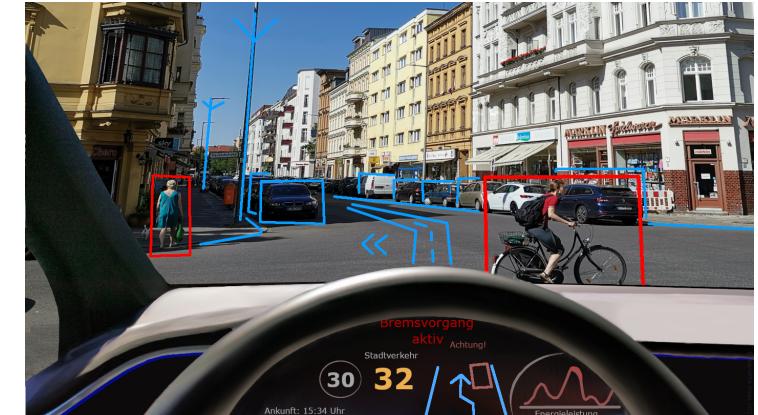
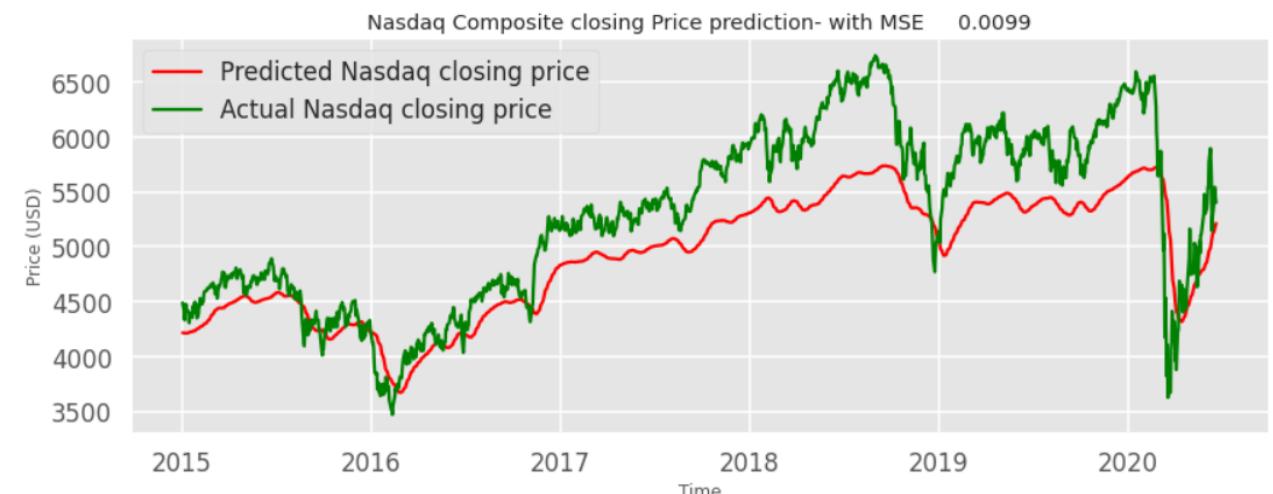
What is Machine Learning?

March 17, 2021

Machine Learning (ML) is a data science technique that learns from existing data to forecast future behaviors, outcomes, and trends

Forms

- **Supervised**
- **Unsupervised**
- Semi-supervised
- Reinforcement/Active



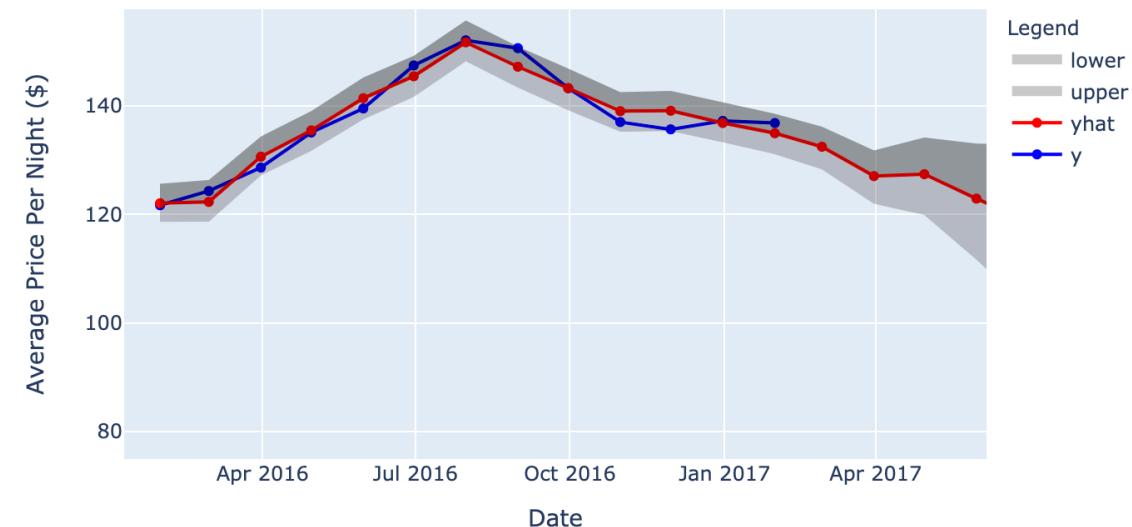
Supervised Learning

March 17, 2021

Supervised ML models are developed by training algorithms on *labeled* data to understand the relationship between the features and the label

- Classification
- **Regression**

Past, Present, and Predicted Future Monthly Average Prices



Forecast Daily Listing Prices

- Modeling using FB's Prophet Library
- Visualization using Plotly

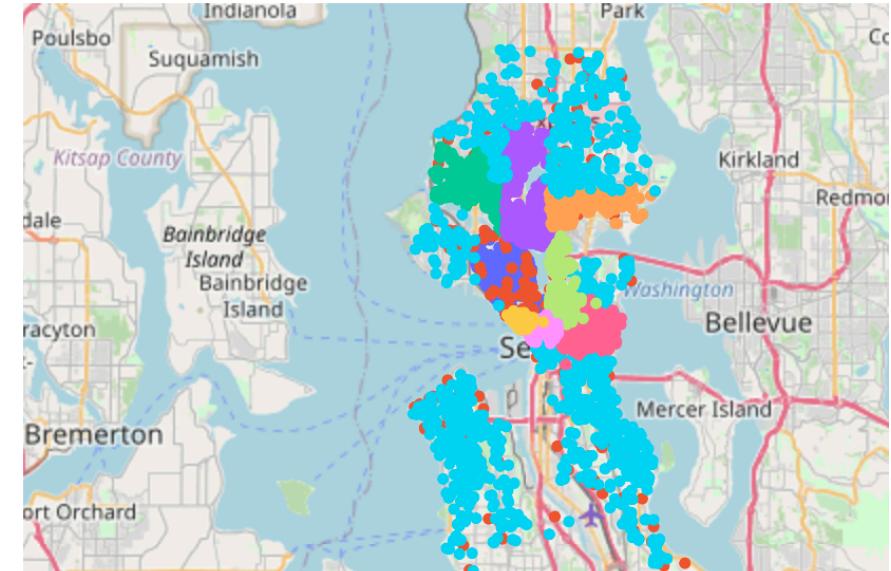
Unsupervised Learning

March 17, 2021

Unsupervised ML models are developed by training algorithms on *unlabeled* data to find relationships amongst variables

- **Clustering**
- Association

K-Means Clustering of AirBnB Spaces



Cluster Listings into Neighborhoods

- Modeling using Scikit-Learn
- Visualization using Plotly

Remember:
a machine
learning model is
designed to
represent the
problem you are
studying, and it is
not a perfect
reflection of the
problem

*All models are wrong
but some are useful*



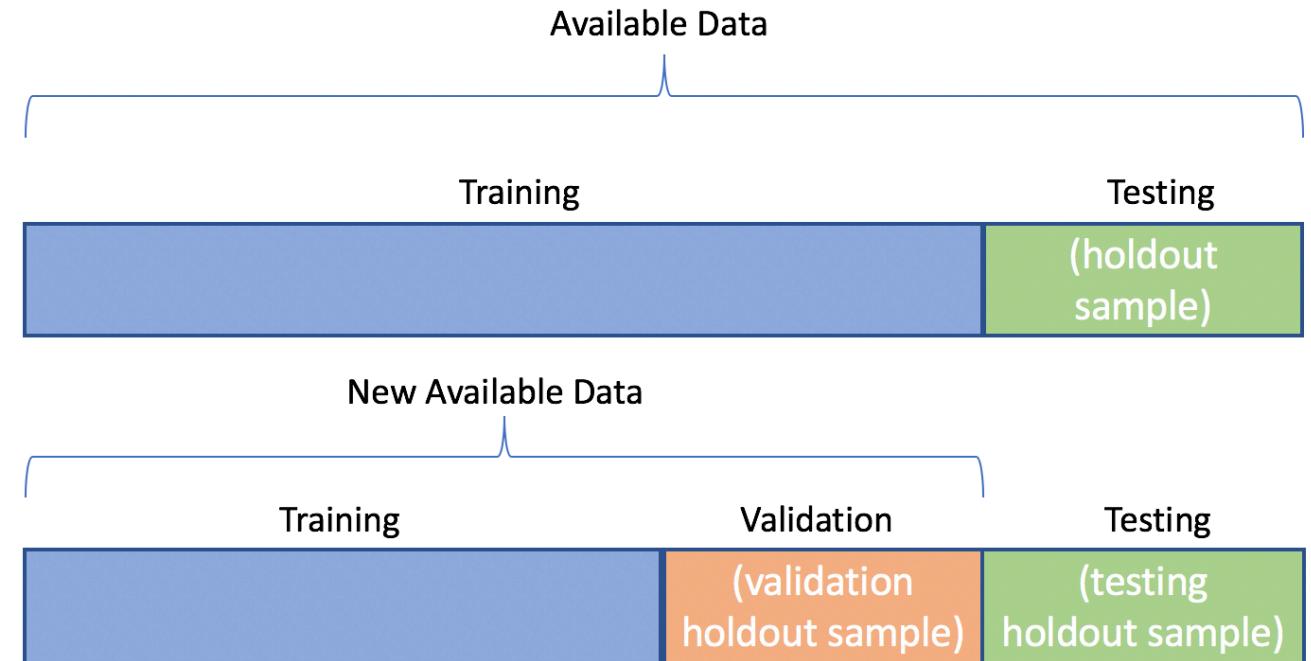
George E.P. Box

Model Training & Evaluation

March 17, 2021

Training & Evaluation (T&E)

- Training: data used to fit (train) the model (typically accounts for 70-90% of all data)
- Evaluation (Test): data used to at the end of fitting to measure how robust the model is to handling other data
- Holdout (Validation): data used to assess the performance of the model after the model has been fitted, tested, and is ready for deployment



High-Level ML Workflow

March 17, 2021

BLUF

- What is the purpose?
- What is the end state goal?
- What are the requirements vs. the requests?

Problem Formation

- What is the goal?
- What are the requirements?

Data Curation

- Collect Data
- Preprocess

Exploratory Analysis

- Learn basic features
- Visualize and discern patterns

Modeling

- Select techniques per Problem Formation
- Train and fit ML pipeline model

Testing & Evaluation

- Iteratively test and evaluate ML model for appropriateness of fit / tune parameters

Integration

- Validate model meets specifications
- Deploy ML tool into production

Considerations

- Product Integration?
- Decision Support Tool or Fully Autonomous?
- Risk Tolerance Requirements 😊



Why Python?





Why Python?

Python is “an interpreted, high-level and general-purpose programming language”

- Versatility
- Code readability
- General and specialized libraries
- Object-oriented design
- **Full data lifecycle support – from Exploratory Analysis to Data Product Development**

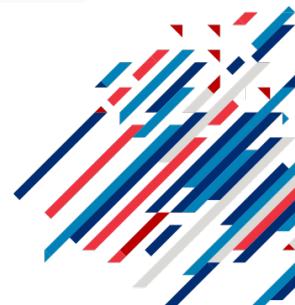
The image shows a comparison between a Python code file and its resulting Streamlit application. On the left, a dark-themed code editor window titled 'MyApp.py' displays the following Python code:

```
import streamlit as st
import pandas as pd

st.write("""
# My first app
Hello *world!*
""")

df = pd.read_csv("my_data.csv")
st.line_chart(df)
```

On the right, a light-themed Streamlit application window titled 'MyApp • Streamlit' shows the output: a title 'My first app' followed by the text 'Hello world!' and a line chart visual.



Python-Powered Applications

March 17, 2021

- Web Development
- Embedded Systems
- Desktop Applications
- Command Line Interfaces (CLI)
- Data Analytics & Machine Learning
 - Curate and model structured and unstructured data
- Scripting



spaCy

Flask



neo4j



Keras



statsmodels



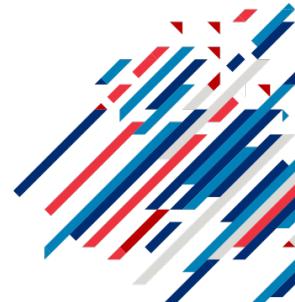
scikit
learn



pandas



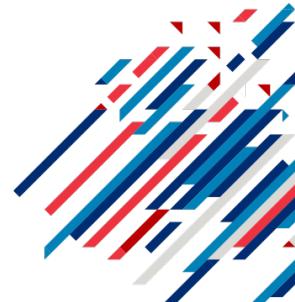
plotly | Dash



Companies that Rely Upon Python

March 17, 2021

- Google
 - "Python has been an important part of Google since the beginning, and remains so as the system grows and evolves. Today dozens of Google engineers use Python, and we're looking for more people with skills in this language." - Peter Norvig, director of search quality at Google, Inc.
- Facebook
- Instagram (a Django app!)
- Spotify
- Netflix (super Jupyter proponents)
- Dropbox
- Pinterest (Django → Flask)
- Goldman Sachs
- Paypal
- Consumer Financial Protection Bureau
- DHS ☺



JOBs

March 17, 2021

Appreciation of Python + MBA is a powerful combo...

- Business Intel Analyst
- Analytics Lead
- Data Scientist
- People Analytics
- Product Manager
- ...

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder

The screenshot shows a LinkedIn search interface with the following details:

- Search bar: python mba
- Location: United States
- Job Alert Off toggle
- Filter buttons: Jobs (selected), Date Posted, Experience Level, Company
- Results count: 1,703 results for "Python mba in United States"
- Job listing 1: **Analytical Lead, Careers & Education** at Microsoft, Washington, DC, posted 2 hours ago. It shows a Microsoft logo and a photo of three people.
- Job listing 2: **Fraud Analyst - Apple Pay - San Diego** at Apple, San Diego, CA, posted 2 days ago. It shows an Apple logo and a photo of one person.
- Job listing 3: **Sr. Business Analyst** at Meritor, Florence, IN. It shows a Meritor logo and a photo of a bull.

Python-Powered ML Tech Stack

March 17, 2021

Analytics

- Scikit-Learn
- Plotly
- Statsmodels, keras, torch, spaCy, etc...

Data Manipulation

- Pandas

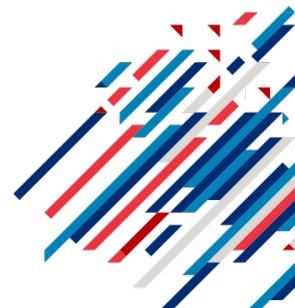
Code & Documentation

- Colab
- Jupyter

Carnegie Mellon University

Tepper School of Business

William Larimer Mellon, Founder





More Business Use Cases



More General Tasks

March 17, 2021

- Clean Data
- Visualize Data
- Predict Labels
- Presence Detection
- Predict Numerical Values
- Detect Topics in Text
- Detect Sentiment
- Summarize Text
- Infer Absolute and Relative Risk
- Find Clusters
- Real-Time Decisions
- Automate Processes
- ...





More Business Use Cases

Demand
Prediction

Optimal Price
Selection

Personalized
Ads

Fraud
Detection

Market
Forecasting

Sentiment
Analysis

Document
Workflow
Automation

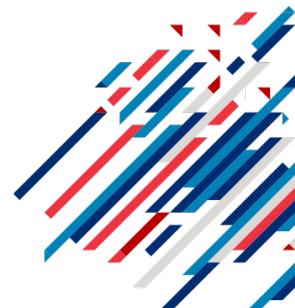
Supplier
Selection

Route
Optimization

Credit Scoring

Algorithmic
Trading

...



Generating Insights for AirBnB

March 17, 2021

BLUF

- Purpose: Clean Data, Visualize Data, Predict Price, Cluster Homes, Parse Reviews

Problem Formation

- Predict Price
- What are the requirements?

Data Curation

- Kaggle AirBnB Data
- Scaling, Imputing, Indicator Variable

Exploratory Analysis

- Plots
- Descriptive Stats

Modeling

- Select techniques per Problem Formation
- Train and fit ML pipeline model

Testing & Evaluation

- Fit Scikit-Learn Model & Statsmodels
- Tuen Parameters

Integration

- Cross-Validate



Let's Apply a Little Python + ML to the AirBnB Dataset with an *Interactive Tutorial*

Time to Code!

GitHub

Github Project Link:

<https://github.com/carriegardner428/ML-with-Python-Tepper-CY21-AW4>

colab

Colab Link:

<https://colab.research.google.com/github/carriegardner428/ML-with-Python-Tepper-CY21-AW4/blob/main/>



Parting Thoughts

- Thank you for your attention!
- Please reach out if you want to...
 - Learn more on these topics
 - Get more hands-on python and analytics experience (e.g., collaborative coding events)
 - Share your knowledge and experience on these topics
- Again: You can do this! 😊

