

# STEPHEN KAPLAN

✉ stevejkaplan@gmail.com  
🌐 blog.stephenjkaplan.com  
☎ 732-927-3873  
📍 Denver, CO 80212  
in kaplanstephen  
🔗 stephenjkaplan

## Skills

### LANGUAGES & FRAMEWORKS

Python  
Flask  
HTML/CSS  
JavaScript

### DATA/ML LIBRARIES (PYTHON)

scikit-learn  
pandas  
XGBoost  
NLTK  
numpy  
Keras  
PyTorch

### VISUALIZATION (PYTHON)

matplotlib  
seaborn  
plotly

### OTHER TECHNICAL SKILLS

Git / GitHub  
Linux Command Line (Bash/CLI)  
Cloud Computing (AWS EC2, GCP)

### MACHINE LEARNING TOPICS

Regression  
Classification  
Natural Language Processing (NLP)  
Recommender Systems  
Unsupervised Learning  
Dimensionality Reduction  
Ensembling  
Tree-Based Models (Random Forest)  
Cross Validation  
Neural Networks  
Computer Vision (Image Classification, Object Detection)  
Boosting  
Time Series / Forecasting

### DATABASES

SQL  
MongoDB

## Education

University of California - Berkeley  
B.S. Mechanical & Nuclear Engineering 2014  
Sigma Pi Fraternity, Iota Chapter

## Experience

### Metis

Data Science Student

Remote  
June 2020 to Sept. 2020

- Completed 12-week, full-time data science bootcamp. Curriculum included machine learning algorithms, software development, statistics, data analysis, data visualization, data engineering and communication.
- Designed and built end-to-end data science projects using a wide range of techniques and algorithms. See "Projects" section below for full details.

### First Solar

Performance & Prediction Engineer

Remote (previously San Francisco, CA)  
Apr. 2016 to Apr. 2020

- Invented and owned [Python SDK](#) product for power plant simulation software "PlantPredict". Enabled 5+ teams across the company, and external partners to automate 20+ processes, reducing time spent on each by up to 95%.
- Researched, prototyped, and validated models to predict annual solar power plant performance within 1% of actual performance for contracted projects in development.
- Presented mismatch loss model at Photovoltaic Performance Modeling Conference (PVPVC) in Freiburg, Germany (2016), and single diode model optimization algorithm at PVPVC in Weihai, China (2018).
- Trained linear regression model using scikit-learn to predict advantage of solar panel over competitor at any location on Earth. Generated geographical heat maps that showed up to 10% performance advantage over competitor, enabling regional technical sales teams to originate 100+ million dollars of solar panel sales.
- Pioneered company's transition to Python and free open source tools from previous MATLAB-based analytics environment.
- Oversaw GitHub issues, sprint planning, code reviews, and education for all internal Python developers.
- Efficiently performed ad hoc analyses and generated reports for various cross-functional internal stakeholders.
- Analyzed solar spectral irradiance data to inform corporate investment in new sensor technology. Queried SQL database, transformed data using pandas, and produced visualizations using plotly to present to internal and external stakeholders.
- Published paper entitled "[Use of the PlantPredict Application Programming Interface for Automating Energy Prediction-Based Analyses](#)" and presented findings at the 45th IEEE PV Specialists Conference.
- Published paper entitled "[Predicting Power Loss Due to Module Mismatch in Utility-Scale Photovoltaic Systems](#)" and presented findings at the 44th IEEE PV Specialists Conference.

### Utility API

Software Engineer

Oakland, CA  
July 2015 to Dec. 2015

- Optimized Python scripts that parse and analyze energy usage data by locating and removing redundant parsing tasks, and grouping similar tasks, increasing the speed of data analysis by greater than 50%.
- Designed monitoring dashboard tools in Python. Reduced data retrieval time from several minutes to instantaneously with nightly data cache.

### BrightBox Technologies

Software Engineer

Berkeley, CA  
June 2014 to July 2015

- Software development of product that optimized HVAC system operation in commercial buildings to save energy, using Python/Django/AngularJS.
- Established Jenkins CI for automated unit test runs before each deployment.
- Designed sequence of operations and commands for automated HVAC system tests and implemented in Python.
- Automated generation of customer-facing report that visualized results of HVAC system commissioning tests.

## Projects

### Spoken Accent Classifier

- Built an ensemble model (composed of a Logistic Regression and K-Nearest Neighbors) using scikit-learn to predict the accent of a speaker from recorded audio. Achieved an accuracy of 89% and ROC AUC score of 0.95.
- Stored data remotely in a SQL database on an AWS EC2 instance.
- Deployed a demo of the model in a [Flask application](#) using Google App Engine.

### Song Popularity Predictor

- Trained and cross-validated a Linear Regression model with Lasso regularization using scikit-learn to predict the popularity of a song on Spotify using data available on their public API. Achieved an R2 score of 0.58 and an RMSE of 13.82 (where popularity is on a scale of 0 to 100).
- Deployed the model with a [streamlit application](#).

### Standup Comedy Recommender

- Scraped transcripts of standup comedy specials using the Python requests library and BeautifulSoup. Stored data in a Mongo database on an AWS EC2 instance.
- Cleaned text corpus using regular expressions, lemmatized using NLTK, and TF-IDF vectorized using scikit-learn.
- Performed topic modeling with NMF decomposition to automatically generate comedy genres.
- Created recommendation system that matches user-defined search terms to selection of standup comedy using cosine similarity.
- Deployed genre filtering and recommendation system [application](#) with Flask.

### Object Detection for Autonomous Snow Grooming Vehicles (In Progress)

- Retraining a pre-trained Faster R-CNN model using PyTorch to detect ski resort-specific objects in video footage.
- Goal is a proof of concept for autonomous snow grooming vehicles.