PGA Tour and Weather Data

PGA ShotLink Data Team

https://github.com/awfuldynne/golf_course_project

Background

- The PGA Tour offers shot level data through the “ShotLink® Intelligence” program
- Does weather have an effect on player performance?
- Quality of weather data provided with ShotLink is lacking
  - AM/PM wind speed, direction
- Dark Sky API provides hourly, historical weather data
  - Both qualitative and quantitative measures
    - Type of precipitation
    - Wind speed/direction
    - Temperature
    - Humidity

I DONT THINK THE HEAVY STUFF IS GONNA COME DOWN
FOR QUITE A WHILE
imgflip.com

Goals

- Provide Python package to help data scientists use ShotLink golf and Dark Sky weather data
    - Simplify the process to retrieve corresponding weather data
- Provide example analyses leveraging shot level data with hourly weather data

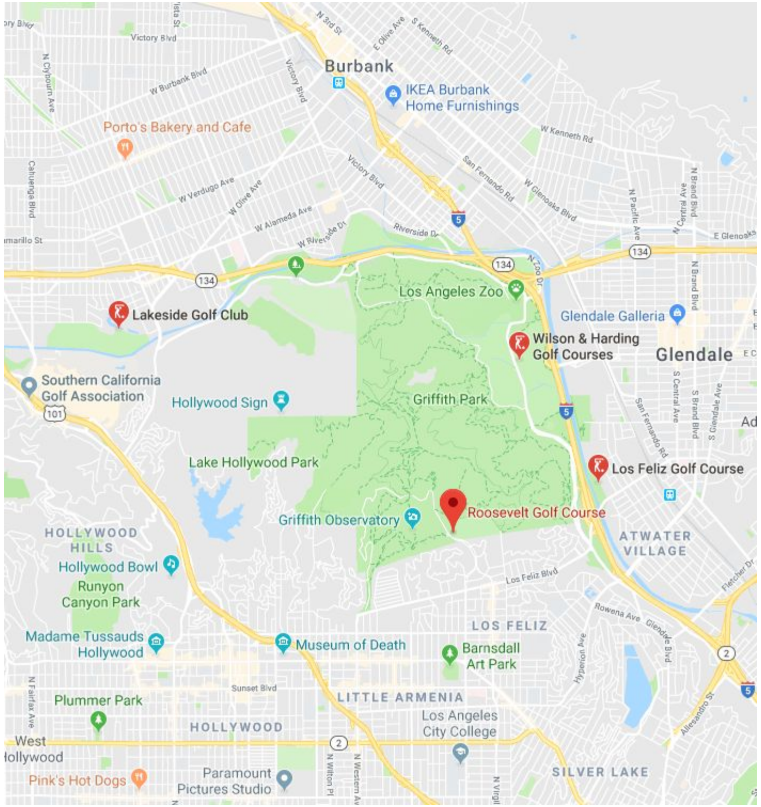
Data Used

- PGA ShotLink
  - **Overview**
    - Collection of shot level and aggregate data from PGA Tour events
    - Focus on Strokes Gained metric
      - How well did a given shot perform based on how the average PGA Tour player plays
  - **Process**
    - Downloaded semicolon delimited files from ShotLink portal
  - **Limitations**
    - Strokes Gained based on 5 year tour average
    - Access to data restricted by application

- Dark Sky API
  - **Overview**
    - Provides an API that allows users to query hourly historical weather data
  - **Process**
    - Used darkskylib python package to make calls to the API
  - **Limitations**
    - Hour is the finest granularity provided
    - Needs latitude and longitude to retrieve data
    - 1,000 free calls per day
      - $1 per 10,000 calls after that

Data Used, continued

- Google Geocoding API
  - **Overview**
    - Maps a given address/location name to latitude and longitude
  - **Process**
    - Python script to iterate over unique course names in ShotLink golf course data set
  - **Limitations**
    - Only have course name from ShotLink data set



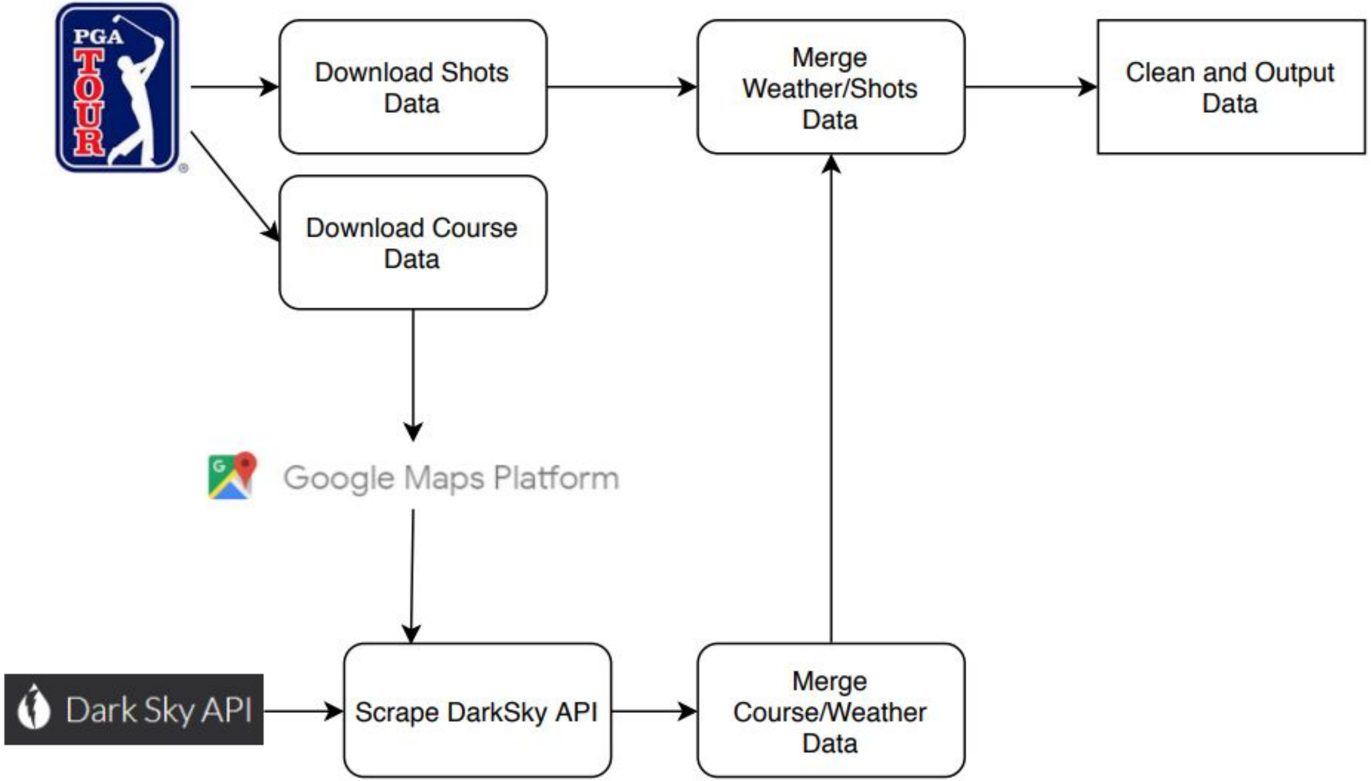
https://github.com/awfuldynne/golf_course_project/blob/master/docs/data.md

Use Cases

- Overall Goal
    - Data scientists can analyze PGA ShotLink data and weather data, together.

- Example Research Questions
    - How does the presence of rain/wind/cloud cover affect golfer performance?
    - How does intensity of rain/wind/cloud cover affect golfer performance?
    - Which golfers are most affected by weather events?

Design Spec



PGA TOUR
Download Shots Data
Merge Weather/Shots Data
Clean and Output Data
Download Course Data
Google Maps Platform
Dark Sky API
Scrape DarkSky API
Merge Course/Weather Data

Github Repo

Project Repository:

- https://github.com/awfuldynne/golf_course_project

Demo - What We Tried

- Visualization:
  - Plot each important weather features vs StrokesGainedBaseline output
- Models
  - Linear Regression
  - Ridge Regression
  - Feature Crosses
- Reprocess the dataset
  - Scale and standardize the training and test data
  - Includes features with NaN values
  - Convert categorical data into indicator variables
  - Upsample rainy days within the training dataset

Demo

YOU'LL GET NOTHING
AND LIKE IT!

Lessons Learned

- Many gotchas to look out for when working with date/time data
- A simple ML model, such as Linear Regression, is probably not sufficient to deduce how weather conditions affect shot performance, especially with such a narrow output
- The overwhelming majority of our shot events took place during nice weather - this makes it difficult to find correlations with the more ‘extreme’ weather conditions; our training set isn’t well-stratified
- The free version of Travis CI (public repo) does work with a private repo dependency, using Git submodules and bespoke encryption/decryption

Future Work

- Further investigation into code coverage solutions
  - Coveralls doesn’t always update with coverage data from Travis
  - But we implemented CodeCov and it seems better
- Player-level analyses
  - Are certain players more effective in rain?
- Deduce wind direction relative to shot direction using wind bearing and sequential shot coordinate values
- Unit test all higher-level ShotLink data cleaning/merging scripts - right now we test all of the core functionality, but only some of the higher-level scripts that call the core functionality
- Try more ML techniques, such as Neural Networks

Questions?