HANK ZHANG

Data Scientist, Software/ML Engineer https://www.linkedin.com/in/hankzhang2003/
hankczhang@gmail.com, (408) 314-6733 https://github.com/hankzhang2003

SKILLS

Programming: Python (+Pandas, Matplotlib, Scikit-learn, Tensorflow, PyTorch), Java, SQL (PostgreSQL, MySQL), R (+plyr/dplyr, ggplot, Shiny), NoSQL (MongoDB), Spark, JavaScript, HTML/CSS

Methodologies: A/B testing, supervised learning, unsupervised learning, classification/regression models, natural language processing, artificial intelligence, Markov chains, neural networks, deep learning

Data management: Tableau, web scraping (BeautifulSoup, Selenium), AWS tools File management: Git, shell scripting, Docker

EXPERIENCE

Galvanize, *Data Science Immersive* Dec 2019-Feb 2020

- 3-month, 600-hour full-time data science intensive program covering core data science/programming topics and skills
- Used programming for data gathering, processing, analysis, and visualization such as Python (Matplotlib, Scikit-learn, TensorFlow, PyTorch), SQL (PostgreSQL), NoSQL (MongoDB), Spark, AWS, and Tableau
- Worked with concepts such as statistics, big data, data processing, database management, web scraping, A/B testing, gradient descent, linear/logistic/regularized regression, decision trees, random forests, gradient boosting, k-means clustering, natural language processing, Markov chains, convolutional/recurrent neural networks, and recommenders
- Worked both individually and collaboratively in projects to demonstrate tools learned
- Developed and presented 3 capstone projects (Spotify Billboard Analysis) and multiple case studies (notable: rideshare churn prediction, IMDB movie review sentiment analysis, movie recommender, fraud detection) using real-world data

DerbySoft, *Data Scientist* Sep 2019-Nov 2019

- Helped manage the database of hotel bookings (6 years, >5m entries per month) using MS Excel, MySQL, and PostgreSQL
- Defined indexes based on OTA metasearch and connectivity data
- Used programming (Python/R/SQL) to query and aggregate hotel reservation and location data
- Exercised machine learning concepts to help create a bid simulator for hotels
- Assisted with significant business insights for data science teams in offices in U.S., China, and Spain

DerbySoft, *Data Analyst* May 2017-Dec 2017

- Used programming (R/Python) to compile, analyze, extrapolate, and simulate metasearch data gathered from Google
- Developed bidding algorithms to solve problems regarding market research, hotel bidding, and decision making
- Helped marketing services and finance departments with connectivity projects and forecasting booking data for 2017
- Presented research to digital marketing managers (DMMs) and discussed possible uses for algorithms and aggregated data
- Contributed to monthly profits and overall 30% growth in the marketing services department

EDUCATION

Northeastern University (Boston, MA) Jan 2018-May 2019

Master's: Analytics, Concentration Applied Machine Intelligence GPA: 3.8

Coursework: Intermediate Analytics, Numerical Analysis, Probability Theory and Statistics, Enterprise Analytics, Linear Modelling, Big Data Management, Data Warehousing, Data Mining Applications, Artificial Intelligence, Information Systems Design

University of California - Berkeley (Berkeley, CA) Aug 2012-Dec 2016

Bachelor's: Applied Math, Concentration Data Science

Coursework: Computer Programs, Data Structures, Machine Structures, Artificial Intelligence, Algorithms, Database Systems

SELECTED PROJECTS

Spotify Billboard Analysis	Analysis of Spotify audio features of the Billboard Hot 100 Songs. Uses exploratory data analysis, classification, regression, natural language processing, and recommenders to visualize and train models for time/feature/lyric data of songs over 60 years. Data gathered using web scraping with BeautifulSoup. Models used/planned were k-means clustering, logistic regression, random forest, gradient boosting, multilayer perceptrons, and convolutional neural networks. Programmed and plotted in Python/SQL, hosted in AWS SageMaker/EC2/S3.
OTA Index Project	Analysis of trends in hotel bookings using Google MetaSearch and OTA connectivity data from DerbySoft. Contains queries and analysis of reservations by country, source countries, destination countries, and destination-source matching. Programmed in Python/R/SQL, plotted in R/Tableau.
Johns Hopkins EBTC ToxCast	Collaborated with Johns Hopkins Evidence-Based Toxicology Collaboration with data analysis. Uses systematic review, in vitro, and human event data to present data pipeline technique and analysis. Cleaned in MS Excel, programmed in R/SQL, plotted in Tableau/R Shiny.
Forecasting	Series of methods to forecast time series data of Google hotel bookings. Contains moving average, exponential smoothing, autocorrelation/residual plots, Holt-Winters/ARIMA forecasting. Programmed and plotted in R.

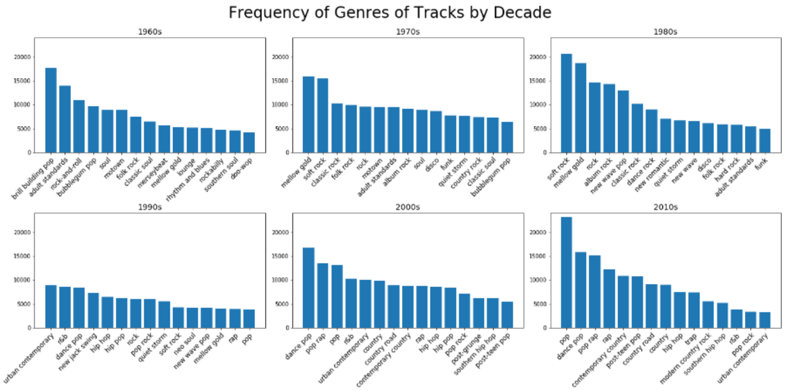
Spotify Billboard Analysis

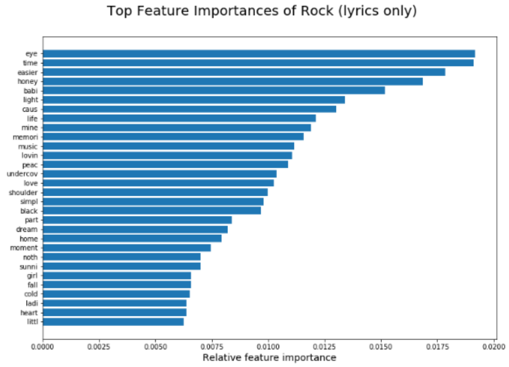
Summary

This is an analysis of the Spotify audio features of songs on the Billboard Hot 100 Songs chart. The objective of this project is to perform data analysis on and make models for the Spotify trends of popular music over the past 60 years. There are many categorical and numerical variables with characteristics of each song, from basic metrics such as title, artist, and album, to metrics that may not even properly expressible with numbers such as energy, danceability, and valence. This project uses exploratory data analysis, classification, regression, natural language processing, and item-based recommenders to visualize and train models for time/feature/lyric data. The data was gathered using web scraping with BeautifulSoup. The models used/planned were k-means clustering, logistic regression, random forest, gradient boosting, multilayer perceptrons, and convolutional neural networks.

Findings

- Features such as energy, danceability, and valence (happiness) are closely related and have higher than average correlation with one another. Likewise, other features such as acousticness and loudness tend to have a negative correlation.
- There is a significant difference between the music features of the songs from the 1960s and the songs of the 2010s. In particular, music today tends to have higher energy, danceability, loudness, valence, and tempo than music of 60 years ago.
- Numerical features can be used to predict the genre of a song with decent accuracy. The top songs of the same genre or genre group tend to have similar features with one another.
- Numerical features usually have much higher impact than words on song happiness. In these models run on this dataset, all of the words combined contributed to less than 5% of what overall determines the happiness of a song.
- In general, it seems like rock songs are easier to predict than pop songs. This is likely due to a higher variety in music styles among pop music, especially in recent years.





Technology Used

Front end: Python (Scikit-Learn, Tensorflow, Matplotlib), AWS (SageMaker), Markdown
Back end: Python (NumPy, Pandas, NLTK, Scikit-Learn, Tensorflow), SQL, BeautifulSoup

Project Link

https://github.com/hankzhang2003/spotify-billboard-analysis