List of Projects

Please click the titles to see details.

Buzz Prediction in Twitter

- Language: R
  - Library: ggplot2, glmnet, randomForest, reshape, gridExtra, dplyr, MASS
- Supervised learning: regression
- Models: LASSO, Ridge, Elastic net, Random forest
- Built 4 machine learning models by cross-validation and evaluated the best one
- Predicted the number of active discussions of a tweet(99% accuracy), paying attention to emergent issues in Twitter
- Identified Top 2 characteristics that most affect buzz tweets(bootstrap)
- Automated variable selection to improve the predictive power

Streaming and Analyzing Yahoo Stock Price in AWS

- Language: Python, Jupyter notebook, SQL
  - Library: pandas, yfinance, boto3, os, subprocess, sys, json, yfinance
- Service: Docker, AWS S3, Kinesis, Athena, Glue, and Lambda function
- Analyzed hourly high stock price by company(streamed Yahoo Finance stock price data into AWS S3 and queried using SQL in AWS Glue and Athena)

Bitcoin Price Time Series Analysis (Deep Learning)

- Language: R
  - Library: keras, tensorflow, ggplot2, dplyr, glmnet, reshape, gridExtra
- Supervised learning: regression
- Models: Deep Learning, LASSO, Ridge, Elastic net
- Developed a program that generates time-series data from history data (using linear algebra)
- Built the best model that predicts the Bitcoin price from past prices by comparing the performance of Deep Learning, LASSO, and Ridge (optimizing them by cross-validation | 97% accuracy)

Analysis on Yelp Business and Customer Patterns

- Language: Python Jupyter Notebook
  - Library: Pandas, PySpark, matplotlib
- Service: AWS S3, EMR
- read over 1 GB dataset from AWS S3 and analyzed in Pyspark(Jupyter Notebook, AWS EMR)
- figured out meaningful business patterns from Yelp reviews(no difference between active and inactive reviewers, cutoff for good service quality, regional characteristics)

Analyzing Millions of NYC Parking Violation

- Language: Python
- Service: AWS EC2, Docker, Elasticsearch, Kibana

- Visualized parking violation patterns of NYC (Top 5 violations, violation trend by time and county, fine reduction | Docker, AWS EC2, Elasticsearch, Kibana)

Nucleus Detection in Cell

- Language: R
  - Library: ggplot2, glmnet, randomForest, reshape, gridExtra, dplyr, MASS, e1071
- Supervised learning: classification
- Models: LASSO, Ridge, Elastic net, Random forest, Logistic regression
- Developed predictive models that classify cell images with a nucleus to facilitate more efficient DNA research (SVM, Random Forest, Logistic, LASSO, and Ridge Regression | 98.5% accuracy | capturing the appearance of a nucleus)
- Applied techniques (oversampling, tuning hyperparameters, dealing with overfitting, regularization | 15% improvement)
- Utilized ggplot library (R) to visually demonstrate which of the models most effectively identifies a nucleus
- Used A/B experiments to reduce the data size while preserving predictive ability (45% reduction in training time)

Parkinson's Disease Diagnosis from Acoustic Features

- Language: R
  - Library: ggplot2, glmnet, randomForest, reshape, gridExtra, dplyr, MASS
- Supervised learning: classification
- Models: LASSO, Ridge, Logistic regression
- Built three binary classification models to identify elderly patients with Parkinson's Disease to facilitate patient diagnoses
- Improved the predictive ability of the model by optimizing hyperparameters (cross-validation | 7% improvement)
- Evaluated the classification models by visualizing the trade-off between model performance and training time
- Figured out important variables that identify Parkinson's Disease efficiently