Lily Liang

612 S Flower St, Los Angeles, CA | 2064839371 | xliang61@usc.edu

EDUCATION

University of Southern California **08/2020-05/2022**
Master of Science, Applied Data Science GPA: 4.0
Relative courses: **Data Management, Machine Learning for Data Science, Data Mining, Algorithm, NLP**

University of Washington **09/2016-06/2020**
Bachelor of Science, Mathematics + Bachelor of Arts, French GPA: 3.64 (Dean's List for 10 quarters)

WORK EXPERIENCE

Data Analysis Intern, Beijing Dataway Horizon Co., Ltd 06/2019-08/2019

- Crawled 7000+ users' evaluations from Tax Service API, cleaned and imputed the data with Python.
- Encapsulated data and computations with **class/object**, generated 20 new features into datasets.
- Summarized statistics, visualized trends by using Python, compared results among different tax offices in Postgre SQL.
- Compiled 7 comprehensive reports for management's review and action, assisted the team to contribute the most in department and distilled findings into representatives shown in the National Tax Services Conference.

PROJECTS

Human Movements' Classification

- Customized data modeling to extract features from multiple files into a pandas dataframe, filled missing data and used **Bootstrap algorithm** to calculate 90% confidence intervals for all features.
- Plotted correlation matrix for some features, applied **Recursive Feature Elimination** in Python, trained **Lasso multinomial logistic regression model**, **Gaussian Naïve Bayes model** and **Multinomial Naïve Bayes model** with **5-fold cross validation**.
- Compared three models by their CV error, test error, **Confusion matrix** and **ROC plot/AUC score**, found Gaussian Naïve Bayes to be the best.
- Utilized **SMOTE** to compensate unbalanced data, trained Gaussian Naïve Bayes model again, noticed CV error dropped by 10% and AUC improved by 5%.

Automatic Statement Generator

- Merged 4 texts files, encoded and partitioned all characters, set the last number of each partition as the class and encoded the class in **one-hot** scheme.
- Built a **LSTM** model with a **single hidden layer** and a **softmax output layer** by using tensorflow.keras, trained the model in 16 epochs to predict further letters from initializing sentences.
- The program generated 1000 more characters to form statements in similar tone to the given sentences.

Bitcoin Prices' Time Series Data Analysis

- Imported time series data into **spark** dataframes, assigned 'Id' to records by their types and repartitioned data based on Ids.
- Reformatted dataframes to train **FB Prophet model**, predicted prices for all old dates and 30 new weekly ones, calculated **95% confidence interval** for all prices.
- Found that all old price records fall into the predicted confidence intervals, plotted prices' trend for all time, designed a user prompter that shows Bitcoin prices by input dates.

Image Colorization with CNNs

- Unpickled and reshaped graphic data into pixels' matrix, assigned each pixel 1 of 4 colors by using **k-means clustering**.
- Utilized color data to train a **deep CNN model** with two sets of **convolution layer** and **MLP layer** by using package keras.
- Obtained greyscale images from predictions and test pixels' colors by using skimage.color package, compared and found predictions highly similar to test images.

Fast-food Popularity and The Correlation to Covid-19 Spreading

- Scraped 20,000+ data from Yelp API and webpages, modeled and integrated data within **pandas dataframes**.
- Created a fast-food popularity measure by using the idea of **Bayesian Averages**, generated a **choropleth map** to present geo-distribution of popularity with folium package.
- Hypothesized and tested the linear correlation between popularity and Covid-19 infect and death data by using **Pearson correlation in degree 1,2 and 3**, analyzed **R-squared scores** and **p-values**.
- Found significant relation between the pandemic and fast-food popularity, composed a report showing the results in detail.

SKILLS

- Programming Language: **Python, R, SAS, Java, Js**
- Tools: **sklearn, tensorflow, MySQL, NoSQL, Spark, Hadoop, AWS**
- Statistical Analysis: **Regression Analysis, Statistical Modeling, A/B testing, Time Series**