Zeyu Yang

(646) 420-3755 | zy2327 @columbia.edu | linkedin.com/in/zeyu-yang | 203 W 108th ST APT 17, New York, NY, 10025

CORE COMPETENCIES

- **Programming Languages**: R (tidyverse, tidytext), Python (Scikit-learn, Pandas, TensorFlow), SQL (Snowflake, MS SQL Server)
- **Models**: Classification, Clustering, Regression, Predictive Modeling, Machine Learning Algorithms (KNN, SVM, K- means)
- **Data Visualization**: R (ggplot2, plotly), Python (seaborn, matplotlib), D3, Chartio

EDUCATION

Columbia University, Graduate School of Arts and Sciences — New York, NY
MA in Statistics, GPA: 3.97/4.0 — Dec 2019

Shanghai University, College of Science — Shanghai
BS in Mathematics and Applied Mathematics, GPA: 3.79/4.0, Rank: Top 10% — Jul 2018

WORK EXPERIENCE

What If Media Group — Fort Lee, NJ
Data Scientist — Jul 2019 – Sep 2019

- Utilized SQL analysis (join, window functions, etc.) on the centralized database, containing more than 1000 tables, to build a new algorithm which updates weekly key statistics to analytics team and saves 90% time to update the summaries than before
- Utilized NLP using python (Pandas, NLTK, spaCy, etc.) to understand the impact of different words & phrases in 100,000 messages, guided push notification team to create attractive contents and increased the Click Through Rate by 30%
- Built a logistics model using python (Scikit-learn) and SQL to reactivate sleeping users based on 1 million user related data and increased the activation rate by 15% compared to the previous model
- Applied data visualization techniques, such as boxplot, bar plot, pie chart, Cleveland dot plot, etc., using ggplot and Chartio, and summarized the overall key statistics (CTR, conversion rate, etc.) trends for technical and market teams to do further improvements

Columbia University — New York, NY
Data Science Analytics — Feb 2019 – May 2019

- Analyzed data for the *Columbia Engineering Graduate Student Quality of Life Survey* to improve the university's facilities and services regarding housing, libraries, dining etc. and to increase students' satisfaction rate
- Identified difference in satisfaction rate among various student groups (program, degree, gender, etc.) and visualized the key findings using ggplot and excel to convey the information accurately and efficiently to all audience
- Added reproducibility to the task by converting the data processing and data visualizations from excel to R
- https://egsc.engineering.columbia.edu/content/quality-life-survey-2019

Edenred China — Shanghai
Data Scientist Intern — Feb 2018 - Jun 2018

- Conducted data analysis for Sephora to help it maintain its Customer Relationship Management in China mainland
- Analyzed customer in-store satisfaction rate regarding shop assistants and store environment from thousands of Sephora Love Meter Survey data and millions of Sephora user data by using SQL and excel and summarized the statistics by regions and stores
- Reduced the time for conducting weekly survey analysis by 75% through converting all the data processing from excel to SQL
- Participated in building a collaborative filtering algorithm to help add commodity recommendation function i.e. recommend products based on similar users for Sephora and increased the unit per transaction by 50%

PROJECTS

Fraud Detection for Online Transaction — Nov 2019 - Dec 2019

- Conducted exploratory data analysis on 600,000 observations and 433 features to check the data quality and discovered that there are a large proportion of missing values in our dataset and the target variables *isFraud* is extremely imbalanced
- Preprocessed the missing values with mean, mode and multiple imputation and applied SMOTE to deal with the imbalanced data
- Experimented on various feature selection method such as PCA, Lasso, Fisher Criterion, etc. and chose Lasso for its high AUC score
- Conducted classification models Logistics Regression and Random forest and achieved AUC score 0.91 using Random forest

Reactivate Sleeping Users — Aug 2019 - Aug 2019

- Identified useful features such as *last_open_days* and *total_opens*, etc. and utilized SQL analysis to join these features from 5 tables that contains user information and click history in centralized database
- Preprocessed 5 million data points, by removing missing values and applied feature selection methods, such as PCA and stepwise methods in the train dataset
- Built logistic regression model as baseline, and further improved the predictive model using XGBoost and Adaboost, as a result, XGBoost returns significant results in accuracy, precision, and F-1 score
- Increase the reactivation rate by 30% than benchmark model

Citi Bike Helper (Shiny App) https://zyang.shinyapps.io/Citi_bike/ — Feb 2019 - Feb 2019

- Led a group of 5 team members to build a public NYC Citi Bike application using Shiny R to help the investors find the best investment location and direct the bike users explore the closest available bike stations
- Scripted the real time NYC bike information in JSON format from Citi bike real-time data in R
- Developed a feature that encourages users to return bikes to a nearby, lack-of-bike station
- Made the app user friendly: presenting the distance and estimated time of the trip; presenting current weather condition