YANG CAI

FIND ME AT:

Telephone: 530-750-9867
Email: yc3404@columbia.edu
LinkedIn: @yang-cai-80165b66
31 River Ct, Jersey City, NJ, 07310
lhttp://bit.ly/yang_cai

COURSES

Experimental design: Advanced data analysis, A/B Testing

Statistics: Probability, Statistical Inference, Bayesian Statistics, Linear regression models

Computing: Statistical Computing, Applied Data Science, Machine learning

DATA ANALYSIS & VISUALIZATION TOOLS

R (tidyverse, ggplot2, plotly, shiny), Python, SQL (PostgreSQL, SQL Server), Tableau

MS OFFICE TOOLS

Excel, Word, PowerPoint

ML SKILLS

Linear regression, logistic regression, KNN, Decision trees, Support Vector Machine, PCA

CERTIFICATIONS

The Complete SQL Bootcamp on Udemy

EDUCATION HISTORY

COLUMBIA UNIVERSITY | NEW YORK, NY
Master of Arts in Statistics | Dec 2018
- Cum GPA: 3.63

NEW YORK UNIVERSITY | NEW YORK, NY
Bachelor of Arts in Economics | May 2017
- Major GPA: 3.46
- Dean's List (Academic year 2015)

RELEVANT EXPERIENCE

TECHNICAL ANALYST INTERN (PYTHON)
New York, NY | September 2019
Crédit Agricole CIB (Loan Review)

- Used **Python** (**pandas, xlwings, openpyxl**), **SQL** and **VBA** to create an automation workflow of Excel reports for analyzing bank's portfolios.
- Created a user interface with **Jupyter Widgets** to help non-technical users generating reports.
- Reduced 90% of the staff-allocation time on the Excel reports task.

NYC YELLOW TAXI (R, TABLEAU)
https://ladyissy.github.io/taxi/ ***Jersey City, NJ | June 2019***

- **Objective:** Predict the likely tip amount for a trip based on the other trip attributes from the NYC taxi data set which contains **3 million** rows.
- Processed and analyzed the complex data set using **dplyr** and visualization tools (**ggplot2**).
- Used filter method for feature selection, built and improved **linear regression** model using Cook's distance.
- Achieved **0.75** R-Squared for the improved model and **0.56** R-Squared for the baseline model.
- Visualized the data set with Tableau dashboard.

PREDITIVE ANALYTICS (R, PYTHON)
New York, NY | November 2018
https://ladyissy.github.io/predictive/

- **Objective:** Work on model evaluation and selection for **predictive analytics** on image data.
- Implemented a baseline model, tried different feature extraction methods such as Canny, diagonal and large neighborhood feature
- Tested different machine learning models: **GBM** and **XGBOOST**
- Achieved an optimal model with 25.23 PSNR and 784 seconds running time for 1500 images.

SCHOOLS HUNTER APP (RSHINY)
New York, NY | October 2018
https://apdatascience.shinyapps.io/school/

- **Objective:** Improve and add business value to an **RShiny** product that previously developed by peers which allows users to explore and compare universities in New York state.
- Expanded the selection of universities into the whole country (US).
- Built an extra function which recommends universities to users based on their selection of certain criteria.
- Redesigned the layout and user interface of the app to make the website more user friendly.

HAPPYDB TEXT MINING PROJECT (R)
New York, NY | September 2018
https://ladyissy.github.io/text/

- **Objective:** Explore HappyDB dataset (a corpus of 100,000 crowd-sourced happy
- moments) to look deeper into causes that make people happy.
- Used word cloud, **natural language processing**, 1-gram frequency, **sentiment analysis** and exploratory data analysis tools (plotly, ggplot2) to find out the differences of reasons that make men and women happy.
- Found out that male gained happiness from playing video games and female did this by talking about their children.