

# SONG TANG

• Email: st883@cornell.edu • Phone: (607) 379-4856 • LinkedIn: <https://www.linkedin.com/in/santostang/>

---

## EDUCATION

---

**Cornell University**, Ithaca, NY

Aug 2017 - May 2018

*MPS in Information Science (focusing on Data Science)*

**GPA: 4.0/4.0**

- Relevant courses: **Machine Learning with Big Messy Data (Python)**, Database Systems(SQL), Data Driven Marketing, Data Visualization(d3.js), Big Data Management (Hadoop), Text Analytics, Design Data Product
- **Good at public speaking:** In MBA presentation class, I was trained to present in the business world ([link](#))

**The Chinese University of Hong Kong**, Hong Kong

Aug 2010 - Dec 2014

*BBA in Finance and Quantitative Marketing*

GPA: 3.6/4.0

---

## WORK EXPERIENCE (2 years in the data science field)

---

**Data Scientist (Advertising)**

Apr 2017 – Jul 2017

*Sagadigits, Hong Kong*

- Developed a data-driven advertising product using **Python** to cluster 200,000 Facebook posts by **LDA** model and obtained 60 groups of interests for each user
  - Used **market basket analysis** to select target audiences that are most associated from advertisers' campaign topic
- Achievement:** Click through rate +5%, Cost per click -20%, Cost per fans -30% by comparing with traditional method of selecting audiences by marketing experience

**Data Analyst (Marketing Science)**

Dec 2014 – Sep 2016

*IBM - The Chinese University of Hong Kong, Hong Kong*

- **Roles:** Delivered **data solutions for 10+ Fortune 500 companies**, analyzed data to find business opportunities, built machine learning models in SPSS/Python/R, presented actionable insight to make business impact
  - **Product Recommendations** for E-commerce store of P&G
    - Used **Hadoop** to perform ETL job on 1TB browsing data to generate behavior features for 500,000 members
    - Used **R** to build market basket analysis to recommend product at right time through right channel
- Achievement:** Activated **46%** inactive customers; transferred **10%** customers from trying samples to purchase
- **Targeted Coupon Prediction Model** for Hyatt Hotel
    - Used **SQL** server to retrieved data of a previous coupon campaign; Cleaned missing values of customer profile
    - Built **logistic regression** in **R** to calculate probability of using coupon; generated customer list to send coupon
- Achievement:** Hit rate +40%, presented actionable insights to the management of Hyatt Hotel

---

## DATA SCIENCE ACHIEVEMENTS

---

- **Book Publication:** *Web Crawling with Python* ([link](#)) by China Machine Press in 2017
  - **Over 10,000 sales** in Alibaba/Amazon/JD platform; collected by over 10 university libraries
  - Wrote the book all by myself
  - 2 year experience of self-learning Python and Web Crawling when working
- Writing an article on **Forbes** with **Lutz Finger (former Data Science director at Snap Inc)** ([link](#))
  - In the course Design Data Product, our team used H1B data to predict the companies that are most likely to sponsor an International student. We were picked by Lutz Finger to write an article about our project on Forbes, **only 3 out of 20 teams** got this opportunity.

---

## DATA SCIENCE PROJECT

---

**Kaggle Competition: Restaurant Visitor Forecasting (models: KNN, GBDT, RF)**

Jan – Feb 2018

- Used **Python** to cleaned and visualized daily visitation data (250K rows) to explore the dataset
  - Generated 80+ engineered features, e.g. lagging visitors (mean, median, max, min) for last 14/28/60/120/180 days
  - Built **KNN, gradient boosting and random forest** using Sklearn and Xgboost to predict visitation for 37 days
    - Selected most important features based on xgb feature importance output
    - Performed cross-validation and grid-search for parameter tuning and model selection
- Achievement:** Ranked **top 22%** with 0.480 RMSE score by ensembling various models

**Box Office Sales Prediction (models: lasso, ridge and huber regression)** ([link](#))

Sep – Dec 2017

- Used **Seaborn** to visualize 1000 movies, analyzed correlation between box office sales and influencing features
- Used **Sklearn** to build Linear Regression models, including lasso, ridge and huber regression to predict box office
- Tuned regularizer through cross-validation, picking up the model with minimum MSE