Yilun (Tom) Zhang

☎ +1 (519) 998 3545
✉ zhangyilun.0719@gmail.com
http://zhangyilun.github.io/

SUMMARY OF QUALIFICATIONS

- Over 5 years experience in statistical analysis, machine learning and data visualization
- Superior capacity to quickly understand and apply new concepts developed through academic and work experience in different industries
- Extensive knowledge and experience in Python (scipy, sklearn, pyspark, tensorflow, keras, matplotlib) , R (tidyverse, ggplot2, dplyr, glmnet, h2o, keras) , and Matlab
- Proficient in various SQL languages and LaTeX; basic knowledge in Java and C++

PROFESSIONAL EXPERIENCE

August 2017 - Now **Data Scientist**, *Manulife Lab of Forward Thinking (LOFT)*, Toronto, Canada.

- Led multiple data science and machine learning projects, built various statistical models and deep learning models (NLP & CNN) to provide insights to business or to semi-automate current repetitive business processes
- Researched on computer vision adversarial attack and defenses, used Grad-CAM to visually understand the effect of attacks on confusing the attention of deep neural networks and effectiveness of available defenses, demonstrated how adversarial attack will potentially influence insurance industries by using transfer learning to generate industry-specific cases, and presented research at Vector Institute ESS #2
- Worked closely with back-end and front-end engineers to design databases, create ETL pipelines for data flow and deploy RESTful APIs for exposing developed machine learning models

May 2015 - Aug 2015 **Data Scientist (Co-op)**, *Athos*, Redwood City, USA.

- Assessed existing algorithm, then individually invented, implemented and tested a new mathematical algorithm to calculate the Athos Score (currently in production app)
- Implemented the first machine learning algorithm to predict lower body heart rate confidence (currently in production app)
- Wrote Python and MySQL scripts to automate data cleaning, transfer and storage processes
- Developed various sensor signal and other data visualization tools for problem detection and algorithm improvement

Sep 2014 - Dec 2014 **Insight Analyst (Co-op)**, *MD Financial Management*, Ottawa, Canada.

- Conducted statistical analysis on a product pre-launch survey; designed and implemented predictive models to identify future product purchasers and a visualization application for the survey data using R with Shiny
- Mined customer and survey databases to explore gender differences in investment behaviour

Jan 2014 - May 2014 **Bioinformatics Data Analyst (Co-op)**, *Princess Margaret Hospital*, Toronto, Canada.

- Implemented an experiment database using Python with SQLite3 which greatly improved efficiency for data standardization and sharing
- Automated experimental data storage to portable SQLite3 database using Python and Excel
- Applied mixed effect models and forecasting on a set of therapeutic experiment data using R; wrote a detailed statistical report in LaTeX
- Analyzed and visualized large scale biological data using Python and R

EDUCATION

Sep 2016 - Aug 2017 **Master of Mathematics, Statistics**, *University of Waterloo*, Waterloo, Canada.

- **Research**: reviewed sentiment analysis literature in depth; implemented semi-supervised machine learning method to classify the sentiment of longitudinal survey final comments by using linguistic variables as well as respondent-specific longitudinal variables

Sep 2011 - Apr 2016 **Bachelor of Mathematics, Statistics and Computational Mathematics Co-op**, *University of Waterloo*, Waterloo, Canada.

- **Research Assistant**: studied the usage of interaction terms in Poisson regression models with particular attention to potential misconceptions within the sociology literature; conducted simulation study to illustrate proper usage of interaction terms in Poisson models in R

Sep 2014 - **Udemy** & **Coursera** & **edX**, *Online Courses*, Remote.

- Data Science Specialization | Machine Learning | Data Visualization (Coursera)
- Introduction to Big Data with Apache Spark | Scalable Machine Learning (edX)
- Python API Development | NLP | CNN | RNN | Unit Testing (Udemy)

OTHER EXPERIENCE | PROJECTS

Nov 2017 **Presented Research at Vector Institute Endless Summer School #2 Representing Manulife**, *Toronto, Canada*.

- **Title**: Adversarial Attacks and Defenses on Computer Vision Systems and Their Impact to Regulated Industries

Apr 2017 - May 2017 **Top 10% & No. 1 in NA, Int'l Data Mining Cup 2017 Representing University of Waterloo**, *Host in Europe, Remote Participation*.

- **Problem**: predict the revenue generated from online shopping actions
- **Challenge**: 3 types of actions (view, basket, order) but only order will generate revenue, however the action type is not given in test data; exact time order of actions are not clear within a day if performed by the same customer (only the date is available); lots of missing values
- **Model**: a single XGBoost regressor to predict revenue directly

Aug 2015 **1st Place, 6sense Data Hack**, *6sense*, San Francisco, USA.

- **Data**: 1.6M reviews, 500k tips by 366k users for 61k businesses from Yelp
- **Solution**: *FoodFates*, an end-to-end web application with map for users to choose potential restaurant location, then it outputs the predicted star rating and existing positive/negative reviews for similar restaurants in nearby area by correlating various characteristics of restaurants
- **Model**: a single Adaptive Boosting regressor for rating prediction; NLTK for sentiment analysis

Feb 2015 **1st Place, Capital One Data Mining Cup**, *University of Waterloo*, Waterloo, Canada.

- **Problem**: predict the break-even bid price for search engine advertisement
- **Challenge**: highly imbalanced dataset; need to calculate/predict multiple intermediate values before calculating the break-even bid (= product revenue $\times$ conversion rate $\times$ approval rate)
- **Model**: a classification and regression tree (CART) to predict conversion rate

Oct 2016 **Code and the City Idea Jam**, *Mississauga, Canada*.

- **Problem**: if Mississauga is planning to extend subway lines, where should the stations be built
- **Solution**: used various open source geospatial data including residence, office buildings, eateries and parking ticket locations, applied clustering algorithm with fine-tuned weighted importance of different data dimensions to identify crowded and vital locations where a subway station will have the most benefits and impact