Jieyi **Deng**

848-252-8487 | dengjy26@gmail.com | Jieyi-Deng | jieyi-deng-

Skills

Programming Skills Python, SAS, R, MySQL, Matlab, PySpark, Scikit-Learn, PySpark

Industrial Knowledge Machine Learning Modeling, ETL Process, Exploratory Data Analysis (EDA), Data Visualization, Hypothesis Testing

Professional Experience

GEICO

Washington, D.C

Modeling Programmer (Data Scientist) *Mar 2019 - Present*

- Responsible for insurance data exploration and analysis on Hadoop cluster; performed feature engineering to provide an explanation version for a business problem and headed supervised learning models in claim predictions.
- Developed customer satisfaction prediction based on valid surveys for 2-year with random forest and XGBoost model with average accuracy 0.72. This project testifies that the expectations toward to insurance adjusters are most important to the overall customer satisfaction.
- Predicted the catastrophic claims for each city in the United States during a hurricane. Executed logistic regression to prejudge the probability of being affect by the storm. Then apply estimate the number of potential claims before the storm landfall.

Waldron Inc.

New York City, NY

Data Scientist - Intern *Jan 2019 - Mar 2019*

- Implemented the recommendation system for an online mobile APP with customized reward strategies for target customers in luxury shopping market.
- Extracted the feature matrix from text description and single item images for the products, e.g. type, occasion, color, and transfer the customer characters, e.g. age, gender, to generate users' feature matrix.

Center for Advanced Infrastructure and Transportation

New Brunswick, NJ

Research Assistant *Sep 2016 - Oct 2019*

- Engaged in data extract, transform and load (ETL) process based on SQL to refine engineering problems; employed feature extraction and visualization to provide data explanation and in-depth comprehension.
- Performed statistical models or statistical testing on R to measure the risk of stochastic events and explained the relationship between accidents and potential accident-causing factors.
- Trained various predictive models (e.g. logistic regression, ensemble methods) based on Jupyter Notebook to recognized the risk pattern and predicted the frequency with an overall accuracy of 0.6.

Projects

Risk Analysis and Prediction of North American Rail Accidents

New Brunswick, NJ

Aug 2017 - Sep 2018

- Collected and combined the pieces of data from various departments and preprocessed it with ETL approach. Then applied EDA to understand the data and find scenarios for further performing the analysis.
- Applied Logistic Regression model as a benchmark and validated model results by Conditional Odds Ratios (COR) to further explain the effects of causing factors on the stochastic accident rates.
- Extracted 30 engineering features and built predictive models using Random Forest, XGBoost and Feedforward Neural Network. The corresponding Receiver Operating Characteristic (ROC) curves are 0.78, 0.87 and 0.85 with 5-fold cross-validation.

Probabilistic Risk Analysis of Flying Ballast Hazard on the High-Speed-Rail (HSR) Lines

New Brunswick, NJ

Sep 2016 - Aug 2017

- Modeled the occurrence of the event as a balanced system and extracted the process into discrete mechanical factors (e.g. wind force, gravity) and mathematically connected each factor with common operating parameters (e.g. train speed, particle density).
- Collected operating data and records from in-site experiments and normalized the statistical distribution of each mechanical factor, then deduced the risk as a cumulative probability.
- Integrated the analyzing process to a flexible Probabilistic Risk Analysis framework and evaluate how the risk increases as 0.019% to 1.2% with the train speeding from 250 km/h to 400 km/h.

Natural Language Processing and Pet Owner Classification on Video Comments

New Brunswick, NJ

Aug 2018

- Preprocessed the comments by removing invalid text and unknown creators, and identified the comments with the label of pet owners or not.
- Extracted features by using tokenizer to split each comment; employed Word2Vec to map each word to a unique fixed-size vector and then transformed each word into a vector using the average of all words in the comment.
- Implemented a Logistic Regression classification to identify the pet owners from comments with regularization parameter grid searching, the area under the ROC curve is 0.91.
- Defined a text analysis function to categorize significant topics to the pet owners by Latent Dirichlet Allocation (LDA) after removing stop words and special characters from the text and the most significant topics are related to training, food and life span.