Yuchen (Winnie) Shao

https://www.linkedin.com/in/yuchenshao/

EDUCATION

Northwestern University Evanston, IL
MS, Analytics Sep 2021 - Expected Dec 2022

University of Washington Seattle, WA
BS, Data Science & Statistics (Applied Computational Mathematics Science) | GPA: 3.8/4.0 Sep 2017 –June 2021

Coursework: Foundation of Computer Science (Java), Data Structure & Parallelism (Java), Database Management (sql), Probability, Intro to Statistics Data Science (machine learning model), Computational methods for data analysis (machine learning, matlab), Resample Interference (bootstrap), Intern data programming, Statistics Software Application (R), Introduction to Statistical Machine learning (R)

SKILLS

Programming skills: Python (Pandas, Numpy, and NLTK), R, Java, SQL
Tools: SQL, Hadoop, AWS, Tableau, Matplotlib, WordCloud, Power BI, Snowflake
Machine Learning: Decision Tree, Random Frost, SVM, Feature Engineering, Logistics regression
Statistics Analysis: Exploratory Data Analysis, survival analysis, longitudinal data analysis, sentiment analysis

WORK EXPERIENCE

Data Science Intern, June 2021 - Aug 2021 **Digital Ocean**, Seattle, WA

- Established an **NLP** classification model to sort customer churn responses into 8 categories with **Python**
- Utilized **SQL** to retrieve customer churn responses, built a program to auto-translate non-English response, and generated word clouds and sentiment distribution on cleaned dataset
- Constructed an **SVM Classifier** with **stochastic gradient descent** optimization method on labeled dataset, implemented 10-fold **cross validation** to find the best set of hyperparameters for the model and increased the accuracy by 8 percent
- Demonstrated the relationship between monthly recognized revenue and different types of churn, and provided insights on future strategy towards churned customers
- Presented the final results to the business operation team to help them understand why customers churned.

Data Analyst Intern, June 2019 – Aug 2019 **Suzhou Academy of Planning & Design**, China

- Attained a guidance of discount strategy to attract more people applying yearly commute cards by visualizing data flow of public transportation population with **Python, Matplotlib, and Tableau**
- Executed data cleaning on 6 million public transportation data of 2018 in Suzhou by consolidating values and removing outliers
- Evaluated and delivered project progress and strategy insights to executive management for transportation discount policy optimization
- Utilized **MySQL** to retrieve card information from transportation table and manipulated data through **Python and R**

DATA PROJECTS EXPERIENCE

Amazon Food Review (NLP, Python) March 2021

- Formed a classification model with **logistics regression** through data cleaning and feature engineering as well as visualized the importance of each feature in the model through seaborn and matplotlib
- Evaluated the model with **ROC, AUC, and confusion matrix** and warehoused the model using pickle package

Conversion Rate Analysis (Python, Jupyter notebook) March 2021

- Determined general user profiles and association between conversion and user profile with exploratory data analysis and discriminated converted and unconverted people through data visualization
- Applied a **random forest** classification model to predict the conversion rate of the e-commerce website.
- Interpreted the model and feature importance, provided general guidance of target audience and ideas to improve conversion rate

Predicting Spam Messages from SMS Message Board(NLP, Python) December 2020

- Conducted exploratory data analysis by using **Numpy, Matplotlib, Pandas, and WordCloud** package in **Python** and employed **NLTK** package to tokenize messages and extract features including word count, URL count, special punctuation count and **TF-IDF**
- Performed forecasting using classifier model such as **Decision Tree Classifier, Naïve Bayes Classifier, and Random Forest Classifier**
- Achieved 0.98 accuracy through directing 5**-fold cross validation** on all Classifier models through **SVM Classifier**

King County House Price Analysis(R) December 2020

- Recognized relation among 15 key features by carrying out data analysis on housing price data
- Forecasted house price by visualization on King County map to stain the house price distribution and potential key features
- Performed **5-fold cross validation** on 14 different best subset model, and identified cross validation error for multiple subsets
- Enhanced the **RMSE** by 0.05 using 5 key features subset while using all the variables as predictors provided RMSE 0.27