Wenjun Sun

Cell: (760)799-0249 || wenjunsun99@gmail.com
LinkedIn: www.linkedin.com/in/wenjunsun

SUMMARY

An undergraduate student who tries to find an internship opportunity on data scientist/engineer positions, with strong programming skills in Python and SQL. Strong backgrounds in machine learning models.

EDUCATION

University of Washington – Seattle, WA **GPA: 3.91/4.0** Expected Graduation: June 2022
B.S. Computer Science, Dean's List
Coursework: Foundations of Computing, Software Design & Implementation, Multivariable Calculus, Linear Algebra

SKILLS

Supervised learning: Random Forest, linear/logistic regression, SVM, Gradient Boosting Machines, Gradient Descent
Feature selection/engineering: one-hot encoding, correlation analysis, PCA, standardization
Tools: **Spark**, Hadoop, Google Collaboratory, scikit-learn, **Python**, **SQL**.

PROJECTS

Built a machine learning system to customer churn rate (GitHub link to notebook)

- Downloaded past customer churn data of 10000 instances and 13 features and loaded into python dataframe using Google Colab.
- Cleaned data using techniques such as dropping unnecessary columns and performing one-hot encoding on categorical variables.
- **Exploratory Data Analysis** and visualization using Matplotlib and Seaborn
- Used sklearn to split dataset into training and testing sets and apply standardization on features.
- Developed ML models on training data using SVM, K Nearest Neighbors, Logistic Regression, and Random Forest, and used **Grid Search** to find the optimal hyperparameters for each model.
- Evaluated best models on various metrics and found that Random Forest is the optimal model with AUC score of 0.854 and recall of 0.479. Business insight is that age is the most important factor in predicting customer churn

Seattle Crime Data Analysis using Spark SQL (GitHub link to notebook)

- Imported 524K records of crime data from 2008 – 2019 from Seattle government website into a Spark dataframe.
- Used **UDF** to extract time of crime occurrence from raw data. Visualization of crime distributions within a day suggests that most crimes happen between 2:00 PM and 11:00 PM and few take places from 12:00 AM to 8:00 AM.
- Discovered that number of crimes increases each year from 2013 (45k crimes/year) to 2019 (51k crimes/year)
- Analyzed the data distribution via **Spark SQL** and found that most crimes happened in Seattle took place in Downtown and Northgate region, and the most frequent crime type is Car Prowl.
- Gained insights that the Seattle police force should focus more in the Downtown areas late at night. Insurance companies should take into the fact that Car Prowl happens a lot in some areas of Seattle.

Movie Recommendation System with Spark (GitHub link to notebook)

- Downloaded 100,000 records of movie ratings about 9,000 movies from 600 users from MovieLens website.
- Split dataset into training/testing dataset. Performed hyperparameters tuning on **ALS recommendation model** with training set. The best regularization parameter is 0.1, with rank of 50. It has RMSE error of 0.897
- RMSE of model is 0.897 on training set and is 0.606 on entire dataset. Therefore, our model did not overfit.
- Used the model to recommend movie "The Salt of the Earth" to user with ID 232 and movie "Three Billboards Outside Ebbing, Missouri" to user with ID 575.
- Learned that "State and Main" is the most similar movie to "The Salt of the Earth" with **cosine similarity** of 0.825.
- Finally, I added my own movie ratings to the training dataset and trained optimal ALS model to recommend movies for my future self. "Coco", "Back to the Future" are the top movies that the engine recommended me to watch.