Andy Seo . Jarvis Consulting

I'm a recent graduate from the University of Toronto with a Bachelor of Science in Statistics, Computer Science, and Mathematics. During my undergrad, I built a strong foundation in Data Science with statistical/mathematical theories, data concepts, data structures, and machine learning algorithms. This field excites me because data allows us to see the problem from a different perspective and help us identify the underlying problem, which we can implement strategies to overcome that problem. Due to my investigative nature, I find great satisfaction when I extract insights from data. I'm very team-oriented and prioritize punctuality over anything.

Skills

Proficient: Python (Pandas, NumPy, Matplotlib/Seaborn), RDBMS/SQL, R, Tableau, Probability/Statistical Modeling (Regression and Classification Models), Jupyter Notebook

Competent: Machine Learning, Data Mining/ Neural Networks, Webscraping, Agile/Scrum, Git, Linux/Bash

Familiar: Hadoop, Java, C, Google Cloud Platform, Virtual Machines, Spark, Databricks

Jarvis Projects

Project source code: https://github.com/jarviscanada/jarvis_data_eng_AndySeo

Cluster Monitor [GitHub]: Helped the Linux Cluster Administration team to monitor node resource usage (e.g.,CPU,memory) in real time. Wrote bash scripts that will help the LCA team easily set up the postgresql instance, create the database and appropriate tables, and save the usage data into the database. Implemented scripts that can run simultaneously on different machines that will track the usages from multiple machines into one database. Used docker to provision the postgresql instance and the automation process was achieved using crontab.

Python Data Analytics [GitHub]: Created a proof of concept (PoC) that will help the LGS marketing by analyzing customer shopping behavior. Used docker to provision the PostgreSQL instance and and loaded the retail data received from LGS into the PSQL data warehouse. Used a SQL client to explore the data/ write queries. Used Jupyter Notebook for analysis, Pandas library for data manipulation, matplotlib for visualizations and numpy for numerical operations. Integrated the RFM Segmentation table that group customers into meaningful categories to help determine the optimal strategies for marketing.

Hadoop [GitHub]: Used big data platforms like Apache Hadoop and evaluated different tools for processing big data. Evaluated Core Hadoop components, including MapReduce, HDFS, and YARN. Provisioned the Hadoop cluster using Google Cloud Platform with 1 master node and 2 worker nodes. Loaded/queried the data to answered business questions using Apache Hive and Zeppelin Notebook. Implemented various strategies to optimize our query execution times and compared their performances to gain a better understanding of what is happening internally.

Spark [GitHub]: Re-implemented the Python Data analytics using Spark and Databricks. Set up Azure Databricks workspace and Spark cluster consisting of 1 master and 1 worker. Big data solution for business problem.

Highlighted Projects

Realtor.ca Analysis for house listings near Richmond Hill [GitHub]: Web scraped realtor.ca near Richmond Hill to obtain 1300 rows of data and proceeded with Data Cleaning, Feature Engineer, Exploratory Data Analysis and created interactive dashboards with Tableau.

Shoe classification using Machine Learning [GitHub]: Differentiate picture of shoe with either a left shoe or right shoe using machine learning. Data gathered and formatted manually, implemented two different versions of convolutional neural networks from scratch using TensorFlow. Plotting the accuracy curves with matplotlib along with automated hyperparameter tuning

Professional Experiences

Data Analyst, Jarvis (2021-present): Automated scripts for node usage information to be stored on the Jarvis database to help the Linux Cluster Administrator team. Analyzed customer shopping trends for London Gift Shop and proposed solutions to increase revenue. Worked with Big data platforms like Hadoop to optimize query times and gain in depth understanding of distributed file systems. Worked with Microsoft Azure Databricks to create workspace and spark cluster to implement big data solutions for LGS.

Education

University of Toronto (2016-2021), Bachelor of Science, Major in Statistics, Minor in Computer Science and Mathematics

Miscellaneous

- Golf
- Tennis