SHANNON MCDONNELL

DATA SCIENTIST

mcdonnell.sc@gmail.com
(716) 264-3393
Sanborn, New York, 14132
shannon-mcdonnell
sconsm

Data Scientist from a background in human sciences, with strong communication skills and business acumen.

SKILLS

PROGRAMMING LANGUAGES

Python
SQL
Bash
PySpark

LIBRARIES

Pandas
Numpy
Scikit-Learn
SciPy
BeautifulSoup
NLTK
Genism
Keras

VISUALIZATIONS

Matplotlib
Seaborn
Tableau
Plotly
Streamlit

MACHINE LEARNING

Regression
Classification
Clustering
Dimensionality Reduction
Natural Language Processing
Neural Networks
Deep Learning
Image Classification

DATA MANAGEMENT

Postgres
SQLite
MySQL
NoSQL
Google BigQuery

EDUCATION

University of Michigan, Ann Arbor — Aug. 2010 to Spring 2014
B.G.S General Studies
Literature, Science and the Arts

Metis - Beginner Python and Math — Aug. 2020 to Oct. 2020

EXPERIENCE

Metis — New York, NY
Data Scientist — Jan. 2021 to Mar. 2021
Completed a full-time, 12-week immersive data science program that provided a uniquely rigorous learning environment.

- Completed rigorous training in python programming in multiple languages (Python, Bash, SQL, PySpark), data wrangling, data engineering, machine learning, statistical modeling, data visualization, project design and communication.
- Designed and built end-to-end projects utilizing core data science tools and algorithms on real world data with applicable results for integration into business environments.
- Produced a portfolio with 5 key data science projects, and presented findings to Senior Data Scientists covering topics of regression, classification, supervised learning, unsupervised learning , machine learning and deep learning.

OrderCircle — San Francisco, CA
Co-Founder and COO — July 2014 to Jan. 2020
OrderCircle is a B2B SaaS company that provides a custom, one-stop-stop to wholesale order management solutions for small-medium sized businesses, across a wide variety of industries.

- Co-founded and operated a small B2B tech startup in the computer software industry that gained designation within the “500 Startups” accelerator in San Francisco, CA.
- Provided executive level team leadership, management and support in all aspects of business development and growth.
- Oversaw multiple small teams of marketers, engineers, and sales personnel focused on increasing wholesale volume for customers via customized online management tools and strategies.
- Implemented high level communication and relationship building to engage new and existing customers and establish, maintain and grow business pipelines.
- Strategically increased sales revenue and business growth through effective problem solving, client partnership management, and business development and sales strategies.

PROJECTS

Scotch Whisky Recommender

Created a Scotch Whisky recommendation system and an interactive web app using Streamlit, that uses NLP and neural networks from online review sources in order to recommend and predict similar Scotch preferences based on user input.

- Created a Scotch recommendation system using NLP machine learning algorithms and deep learning neural networks to discover flavors, notes, and finishes of Scotch Whisky.
- Utilized NMF topic modeling, logistic regression for predictive classification using TF-IDF vectorization and Word2Vec embeddings, and LSTM neural networks with GloVe implementation for transfer learning.
- Performed predictive analyses and topic modeling on all raw text from online review sources scrapped from WhiskyAdvocate.com, America's leading Whisky magazine (gathered from Kaggle dataset by thatdataanalyst).
- Implemented findings and results on an interactive web app, using Streamlit, that allows users to filter and find the perfect Scotch Whisky. The user can visualize and interact with plotly graphics displaying some of the pertinent EDA to help them make their input decisions, which will produce a recommendation based on pairwise distances using cosine similarity.

Analyzing Children's Stories

Used NLP to analyze the importance of children's stories, and how NLP is especially useful in the early childhood development space. Using data gathered from 'Project Gutenberg', the objective was to find out why it's so important to read to our children and what overall themes are present in children's books, tales and fables.

- Performed multiple preprocessing techniques on the raw text of over 120 children's stories, and executed NMF and LDA topic modeling using both count and TF-IDF vectorizations to find the best overall topics of the children's books dataset.
- Derived the final topics from the NMF model using TF-IDF vectorization: nature and outdoors, animals, mythical creatures, fairy tales, and family and relationships.
- Produced multiple visualizations using matplotlib, including wordclouds and word frequency distributions among top authors and across all major topic categories.
- Referenced scholarly research on the importance of reading in early childhood development. Utilizing domain knowledge to connect the importance of NLP in childhood texts and the impact that these stories have on growth, understanding and development.

Classifying Chicago Restaurant Inspections

Used classification modeling to predict whether or not a restaurant in the city of Chicago will pass or fail inspection. Utilized machine learning techniques to help improve restaurant quality in the Chicago area, as well as help prevent the spread of food borne diseases. Data from over 200,000 restaurants was gathered from the City of Chicago's online open source data portal.

- Performed multiple variations of KNN classification modeling and hyper-parameter tuning, in order to best predict restaurant inspection results.
- Analyzed multiple evaluation metrics including accuracy, precision, recall, F1 scores, and ROC AUC. Ultimately, deciding on the F1 Score as the main decision metric.
- Created a Tableau mapping of the geographical features where certain levels of risk are higher or lower depending on location.
- Concluded that areas closer inland, as opposed to near the lake, are predicted to have a higher likelihood of passing inspections. Ongoing work consists of comparing zip code locations across median household incomes, and looking at specific examples of restaurant types and cuisines in order to add depth to the data and improve our prediction metric.

Predicting IMDB Total Movie Grosses

Built a regression model to predict total movie theater grosses for domestic movies, by web scrapping data from IMDB's Box Office Mojo using BeautifulSoup. Performed multiple regression modelings, focused on evaluation metrics of R2, MSE and MAE values.

- Scraped roughly 3,000 domestic movies on Box Office Mojo using BeautifulSoup, and performed feature engineering on the data, using matplotlib and Seaborn for visualizations.
- Performed feature scaling and normalization, and evaluated all regression models: linear regression, LASSO regression, ridge regression, and ridge cross validation across all metrics.
- Concluded the best predictive results to maximize total movie grosses was a simple linear regression model. The final features used in getting the best predictive model were monthly movie gross, the number of theaters the movie was released in, the release month, and the release year. This final model was evaluated based upon on the results of R2, MSE and MAE values.

Exploring NYC Subway Usage

Explored NYC Subway MTA turnstile data in order to discover the most optimal subway stations and street team placements for the promotional activities of our client. Used python programming and EDA to discover the best possible locations that would maximize the charity's goals to increase awareness, fill event space, and raise funds.

- Collaborated with a team of data scientists in order to come up with the best and most creative solution to the client's problem.
- Gathered several datasets for this analysis, from the publicly available MTA data, NYC housing sales data, and US Census data.
- Located the top NYC subway stations with the highest average foot traffic, analyzed the the most optimal days of the week, as well as targeted wealth and gender of specific areas for supplemental information. Concluded with a detailed analysis of our findings, as well as provided detailed score charts and visualizations to the client.