Samuel Ramirez

Portfolio

(718) 664-8944
samuel.ramirez21@my.stjohns.edu
South Ozone Park, NY
https://www.linkedin.com/in/samuel-ramirez-aa674a269/
https://github.com/samuelramirez21/Projects

See more

Check out my LinkedIn profile: https://www.linkedin.com/in/samuel-ramirez-aa674a269/

My Profile

ABOUT ME

My name is Samuel Ramirez and I am a problem solver who is hardworking, detail-oriented, and has a passion for the fields of Actuarial Science and Data Analytics.

I am a highly motivated professional possessing dual master's degrees in Actuarial Science and Data Science at St. John's University, building on a strong foundation from my Bachelor of Science in Mathematics (with minors in Statistics and Economics) from the University at Buffalo. My experience spans roles in both the public and private sectors—from streamlining fuel management compliance as a Fleet Aide/Analyst at NYC DCAS, to optimizing financial data and IT operations at Cicatelli Associates Inc., to handling community-focused tasks at NYC Parks & Recreation. Alongside passing several actuarial exams and earning an AWS Cloud Practitioner certification, I have developed proficiency in SQL, Python, R, Scala, and Java, as well as working with big data tools like Spark, Hadoop, and Hive. With a passion for transforming complex data into actionable insights, I thrive on problem-solving, collaboration, and leveraging technology to drive meaningful results.

Chain-Ladder vs. Bornhuetter-Ferguson: Worker's Comp

https://github.com/samuelramirez21/Projects/blob/main/Chain-Ladder%20vs.%20Bornhuetter-Ferguson%20-%20Workers'%20Compensation.pdf

This project compares two key actuarial reserving methods—Chain-Ladder and Bornhuetter-Ferguson—to determine which produces more reliable forecasts for the Workers' Compensation line of business. Using both paid and incurred loss data from 57 insurers, several loss development factor (LDF) models were applied, and results were evaluated through Mean Absolute Error (MAE). Overall, Bornhuetter-Ferguson demonstrated slightly better accuracy due to its stability and blend of expected and development techniques. However, the Chain-Ladder method still performed better for a significant subset of the insurers, and thus both methods are recommended in practice to account for varying data characteristics.

SKILLS

Excel
SQL
Python
R
Tableau
PowerBI
Spark
Hadoop
HDFS
Hive
Linux
PowerShell
AWS
Java
Scala
MS Word
PowerPoint

Music Rank Project

https://github.com/samuelramirez21/Projects/blob/main/Music%20Rank%20Project.pdf

This project explores what features help propel a song to popularity on major music streaming platforms—Spotify, Apple Music, Deezer, and Shazam—by analyzing a dataset of 2023's top-charting tracks. Through techniques such as linear regression, best subset selection, cross-validation, decision trees, random forests, boosting, ridge, lasso, and principal component analysis, the study identifies key predictors (like speechiness, acousticness, and energy) that play a role in a song's ranking, while also highlighting variables (like mode and key) that appear less influential. Although the best models explain only around 4% of the ranking variance—indicating that many factors outside of the dataset likely influence a song's popularity—this work underscores the intricate, multifaceted nature of music trends and reveals new insights into how intrinsic song characteristics contribute to chart success.

Stock Trading Strategy

https://github.com/samuelramirez21/Projects/blob/main/Stock%20Trading%20Strategy.ipynb

In this project, I built a diversified stock portfolio across three sectors—Consumer Services, Finance, and Energy Minerals—and then identified an optimal short- and long-term simple moving average strategy for each sector. By testing various moving average combinations (e.g., 5-day vs 200-day) on 2020 data, I selected the combination that maximized gains while minimizing frequent trades. Afterward, I applied the best strategy from each sector to historical data from early 2021 through late 2022. The Energy Minerals sector strongly outperformed the other two, offsetting their losses and resulting in an overall portfolio profit of more than $22,000 on a $100,000 initial investment. This outcome highlights two major takeaways: first, the importance of diversification in mitigating losses when certain sectors underperform; and second, the value of tailoring moving average crossovers to a stock's sector-specific characteristics to optimize returns.

Analyzing Gender Disparity

https://github.com/samuelramirez21/Projects/blob/main/Analyzing%20Gender%20Disparity.ipynb

This project analyzes a global dataset from various STEM fields—spanning genders, ages, job roles, and countries—to understand how compensation is influenced by social and demographic factors. The findings reveal a clear gender imbalance, with males not only comprising the majority of the workforce but also earning higher median salaries than females and nonbinary participants. Age shows a weak positive correlation with salary, indicating that income tends to increase with experience, though it begins to level off and even dip slightly for workers beyond their 40s. Among job titles, Operations Research Practitioner commands the highest median salary, whereas Programmer ranks lowest. Additionally, significant differences emerge across countries, with the United States and Switzerland boasting the highest median salaries and Nigeria and Egypt at the lower end of the pay scale. Taken together, these disparities underscore the need for targeted efforts to address equity issues in compensation and representation within the global STEM community.

PERSONAL

During my undergraduate years in the University at Buffalo I was an active member of Intervarsity Christian Fellowship where I enjoyed engaging with the student population and gauging interest in chapter membership while also attending winter and summer conferences. I was also an active member of the Pokémon Club where I was nominated and voted in as secretary for the 2017-2018 academic year and attended weekly meetings where we would participate in Pokémon tournaments. One of my hobbies include attending the gym with a focus on sustaining both my physical and mental health. I also enjoy going to the theater with my friends and keeping up with the Marvel Cinematic Universe. I am also a fan of Game of Thrones and have thoroughly enjoyed its latest installment: House of Dragons.

Data Mining Competition

https://github.com/samuelramirez21/Projects/blob/main/Data%20Mining%20Competition.ipynb

I competed in a Kaggle challenge to build the most accurate classification model and secured first place with an impressive score of 0.98563. Using Python's data science stack (pandas, NumPy, scikit-learn), I began by performing feature scaling and principal component analysis (PCA) to reduce the dataset from 988 features down to a more manageable 74. This step helped mitigate overfitting and boosted model efficiency. Next, I trained a Multi-Layer Perceptron (MLP) Classifier, fine-tuning parameters to maximize performance. To validate the robustness of my approach, I applied 10-fold cross-validation, which consistently yielded high accuracy (around 98%). Finally, my model's predictions on the unseen test data achieved the highest accuracy score on the leaderboard, demonstrating both the effectiveness of dimensionality reduction techniques and a well-tuned neural network.

Generating Word Embeddings

https://github.com/samuelramirez21/Projects/blob/main/Word%20Embeddings.ipynb

This project showcases the development of high-dimensional word embeddings using the Word2Vec algorithm on a large collection of PubMed abstracts. First, a corpus of 132,935 medical abstracts is transformed to lowercase and accented characters are removed to ensure consistent processing. Each abstract is then tokenized, stripped of punctuation and stopwords, and lemmatized to normalize word forms. Using both the Skip-Gram and CBOW architectures, eight total Word2Vec models are trained with different window sizes (2, 5, 10, and 20) and an embedding dimension of 2048. The resulting embeddings are tested through illustrative "most similar" queries, demonstrating their ability to capture semantic relationships (e.g., "obese + healthy – sick"). Finally, each trained set of word embeddings is exported as a ".emb" file, making these models ready for integration into downstream natural language processing tasks.

Data Mining in Videogame Sales Dataset

https://github.com/samuelramirez21/Projects/blob/main/Data%20Mining%20in%20Videogame%20Sales.ipynb

This project explores a large dataset of video game sales from 1980 to 2020, employing a variety of data mining techniques to uncover insights. After data preprocessing—such as cleaning missing values, removing duplicates, and encoding categorical attributes—exploratory data analysis and visualizations highlighted sales trends across different regions, platforms, genres, and publishers. A decision tree classifier revealed that "Genre" is the best nominal predictor of high sales, while K-means clustering on the numerical sales attributes partitioned games into three "success-level" clusters—moderate, high, and critical. Finally, applying the Apriori algorithm exposed hidden relationships between publishers, genres, and platforms, showing, for instance, that certain companies specialize heavily in particular genres or specific platforms. These combined insights illustrate both how publishers like Nintendo and Sony vary in their strategies and how certain titles become industry-defining critical successes.

Titanic Dashboard

https://github.com/samuelramirez21/Projects/blob/main/Titanic%20Dashboard.pdf

I developed an interactive Titanic Passenger Dashboard by combining six distinct data visualizations in Excel to uncover insights about survival rates, age/sex breakdowns, passenger classes, ports of embarkation, and more. In the process, I utilized a range of Excel's core functionalities—including pivot tables, sorting and filtering, IF formulas, and VLOOKUP—to clean, analyze, and link the underlying data. The result is a cohesive, visually compelling summary of Titanic statistics that demonstrates my ability to translate raw data into digestible, action-oriented information.

Pokémon Dashboard

https://github.com/samuelramirez21/Projects/blob/main/Pokemon%20Dashboard.pdf

I created a comprehensive Pokémon Dashboard in Tableau that highlights key battle statistics across various Pokémon types. By aggregating and visualizing offensive, defensive, speed, and overall strength data, this dashboard reveals that Ground-type Pokémon boast the highest offensive stats, Steel-types excel defensively, Flying-types are the quickest, and Dragon-types lead in overall battle strength. The combination of bar charts, treemaps, and color-coded comparisons provides an intuitive way to explore these insights, demonstrating my ability to derive meaningful conclusions and effectively communicate them through interactive data visualizations.