Sahil Nanavaty

M.S. in Applied Data Science | Syracuse University

Data Science Portfolio

Milestone report

**Introduction**In the pursuit of my Master of Science in Applied Data Science from Syracuse University, I have engaged in a variety of projects that not only solidified my theoretical knowledge but also enhanced my practical skills in data science. This portfolio represents a curated collection of projects that I undertook during my coursework, each chosen for its relevance to the core learning outcomes of the program. These projects illustrate my ability to apply data science techniques across different contexts and showcase my evolution from a student to a data science practitioner ready to tackle real-world challenges.

Each project selected for this portfolio addresses one or more aspects of data science, from data collection and management to advanced analytics and visualization. Through these projects, I have demonstrated my ability to synthesize information, develop strategic insights, and communicate findings effectively, all of which are critical competencies for a data scientist.

This paper will analyze each project in detail, linking the practical work completed to the theoretical concepts learned and the program’s learning outcomes. This reflection not only shows my individual accomplishments but also illustrates how each course has contributed to a comprehensive skill set that prepares me for my future career.

**Project 1: Steam Player Count Analysis – Core Data Science**

In this project, I aimed to analyze the player count data from the Steam digital distribution platform to gain a comprehensive understanding of trends within the PC gaming market. The primary objective was to determine which games were most popular based on average and peak player counts, as well as monthly gains and losses. Furthermore, I sought to understand the timeliness of these player counts to ascertain if there were specific times of the year more favorable for game releases or updates. This analysis is essential for strategic decision-making concerning the scheduling of marketing and publishing games.

The project involved the use of R to meticulously gather and organize extensive player data sourced from Kaggle, ensuring a robust dataset for analysis. Through the application of statistical techniques, I identified patterns and trends in player engagement across different games and throughout various seasons. Additionally, I developed scripts in RStudio to automate the processing and analysis of data. This not only enhanced the efficiency of the data handling but also improved the interpretability of the results. The project required a comprehensive understanding of data management and computational scripting, both of which are integral to the analysis of large datasets.

This project exemplifies my ability to tackle a real-world business problem using appropriate data science techniques. Through the management and analysis of large datasets, I have demonstrated foundational skills crucial for data-driven decision-making in any business context. The Steam Player Count Analysis project showcases my competence in several key areas: identifying business problems, determining the data required to address these problems, and applying suitable analytical techniques to derive meaningful insights that can guide business strategies.

Tools:

**R:** Language used for statistical analysis and data manipulation

**RStudio:** Development environment for R

**dplyr:** Package for data transformation

**tidyr:** Package for organizing raw data

**readr:** Package for importing large datasets

**lubridate:** Package for analyzing date-time data

**ggplot2:** Package for creating visualizations

**Project 2: DataQuest Account Administration – Database Management**

For this database management project, I aimed to design and implement a database that could efficiently store and organize membership data for a fictional massively multiplayer online game (MMO) titled DataQuest. The primary goal was to develop a robust database system that facilitated the management of user accounts, characters, and guild memberships – a real-world application essential for gaming platforms.

I began by defining the stakeholders and business rules crucial to the database structure. Stakeholders included DataQuest publishers, developers, and players, each with distinct needs and interactions with the database. The business rules specified the relationships between accounts, characters, and guilds, including account tiers and guild memberships.

To align with the project objectives:

* **Conceptual and Logical Modeling:** I developed both conceptual and logical data models to outline the structure of the database and the relationships between its entities. This step was crucial in ensuring that the database would effectively support the required operations.
* **Data Definition Language (DDL) Statements:** Using SQL, I created DDL scripts to establish tables for accounts, characters, guilds, and character-guild associations. These tables were designed with primary keys, unique constraints, and appropriate foreign key constraints to maintain data integrity and support complex queries.
* **Normalization:** I applied database normalization principles to ensure that the database structure was optimized for query efficiency and to minimize redundancy. This process was essential for maintaining data integrity and supporting the database's scalability.
* **Data Insertion:** I populated the database with realistic data entries to reflect typical user interactions and to facilitate realistic scenario testing.
* **Complex Queries:** I developed complex SQL queries to answer specific data questions, such as identifying free-to-play users, average age of players, and players not belonging to any guilds. These queries demonstrated the database's ability to provide actionable insights into player behavior and system usage.
* **View Creation:** To streamline data access and reporting, I created SQL views that simplified the retrieval of common data sets, such as contact information for all players and details about free-to-play players.

In addition to these steps, I ensured the database design adhered to principles of ethics and privacy. This involved implementing strict access controls and data encryption to protect sensitive user information, maintaining transparency about data usage policies, and ensuring compliance with relevant data protection regulations.

I developed database procedures, functions, and packages to process the data and implement business logic. I then rigorously tested the database to ensure it met all functional requirements without errors. Optimization techniques were applied to enhance query performance and ensure scalability. This included indexing critical columns and refining SQL queries for efficiency.

The project was an invaluable learning experience that sharpened my skills in database design, SQL programming, and database management. It challenged me to apply theoretical knowledge to a practical problem, deepening my understanding of relational database systems in a real-world application context. My successful navigation of the project's complexities – from initial design through implementation – underscored my competence in handling advanced database tasks and prepared me for further professional challenges in data-intensive environments. The final report included comprehensive documentation of the database schema, SQL scripts, query results, and a thorough analysis of data query outcomes.

Tools:

**SQL:** Language used for creating/manipulating database structures and querying data

**MySQL:** Relational database management system

**Project 3: Loan Analysis – Business Analytics**

The Loan Analysis project delves into the factors influencing whether a customer decides to take out a loan, shifting the focus from general assumptions to data-driven insights in a business context. By employing Logit and Probit regression analyses, I identified significant predictors among various customer attributes such as monthly average spending, education, family, and income, while excluding customer IDs and ZIP codes. Initial results underscored the importance of certificate of deposit ownership and credit card ownership, revealing that having a certificate of deposit increases the likelihood of taking a loan, whereas having a credit card decreases it. This data-driven approach provided a solid foundation for understanding customer loan behaviors.

To deepen my insights, I explored interaction effects between key variables, such as family and income, family and education, monthly average spending and income, and age and income. The interaction between family and education stood out, showing that higher education combined with a larger family size decreases the likelihood of taking a loan. This finding aligns with the notion that higher education typically correlates with higher income, reducing the need for additional loans. Similarly, the interaction between monthly average spending and income indicated that higher spending alongside higher income increases loan likelihood. These interaction analyses added layers of understanding to the initial findings, highlighting the nuanced relationships between customer attributes and loan-taking behavior.

In the final model, I focused on variables that proved both statistically significant and conceptually relevant: monthly average spending, education, family, income, and the interaction between family and education. Additionally, I conducted a neural network analysis to compare these traditional models, revealing that average monthly spending and income had a more substantial impact on loan decisions than education and family size. This project highlights the value of creating actionable insights using the full data science lifecycle. By leveraging Microsoft Excel for both statistical analysis and sensitivity training, I provided valuable, data-driven insights into customer behavior, aiding financial institutions in making informed, data-driven loan decisions. This comprehensive analysis ensured robust findings that could be communicated effectively to both technical teams and stakeholders, demonstrating the practical application of data science methodologies in real-world scenarios.

Tools:

**Microsoft Excel:** Software used for Logit and Probit analysis

**Project 4: Movie Recommendation System – Text Mining**

The proliferation of streaming services has drastically changed how audiences choose what movies to watch. Gone are the days when decisions were primarily influenced by word of mouth; today, viewers are presented with an overwhelming number of choices across various platforms.   
To address the plethora of choices on digital streaming platforms, I lead the development of a movie recommendation project to increase user engagement by tailoring movie selections to individual preferences. This project is a vital application of data science, demonstrating how predictive analytics can be utilized to enhance user engagement through personalized content delivery.

To tackle this challenge, I implemented several machine learning models, each chosen for their strengths in handling different types of data and prediction tasks:

1. **Naive Bayes classifiers**: Utilized for their efficiency in making predictions based on the probability of an event, particularly useful for large datasets of movie descriptions.
2. **Support Vector Machines (SVM):** Employed for their robustness and effectiveness in high-dimensional spaces, which is ideal for text classification tasks involved in analyzing movie descriptions.
3. **Clustering algorithms:** Applied to segment user preferences into distinct groups, facilitating targeted recommendations.

For preprocessing, I used natural language processing techniques, including TF-IDF vectorization, to transform text data into a format suitable for machine learning models. This was crucial in handling the vast datasets of movie descriptions and reviews.

By deploying these models, I successfully generated actionable insights that refined the movie recommendation process, directly enhancing the user experience and increasing user engagement. This required a deep understanding of each model’s strengths and the ability to critically analyze their performance using metrics such as accuracy, precision, and recall. The project highlighted my proficiency in applying complex data science methodologies, including predictive analytics and unsupervised learning, to deliver tangible outcomes in a real-world application.

Tools:

**Python:** Language used for implementing machine learning algorithms and data manipulation

**Natural Language Toolkit (NLTK):** Set of libraries for symbolic and statistical natural language processing (NLP)

**Scikit-learn:** Library for machine learning that provides tools for data mining and analysis

**Pandas:** Library for data manipulation and preprocessing

**NumPy:** Library for handling large, multi-dimensional arrays and matrices

**Matplotlib:** Library for creating visualizations

**Seaborn:** Library for creating visualizations

**TF-IDF Vectorizer:** Tool for converting text data into a matrix of TF-IDF features

**Project 5: Disney Movie Success Story – Data Visualization**

The Disney Movie Success Story project involved analyzing historical box office data from Disney movies to understand the factors that influence their success in terms of box office performance. The aim was to identify trends and patterns that could assist producers in predicting future successes and replicating the studio's past achievements. This required a deep dive into data visualization techniques to ensure the communication of insights was both effective and intuitive.

In this project, I employed R for data cleansing and preparation, ensuring that the data was accurate and reliable for analysis. I developed multiple visualizations to identify distributions and relationships using R plots, focusing on clarity and interpretability to effectively communicate findings to stakeholders with varied levels of technical expertise. Adobe Illustrator was used to enhance these visualizations, combining data plots, design elements, and contextual cues into a single coherent artifact.

This project showcases my skill in using data visualization tools to communicate complex data insights in a clear and effective manner. It demonstrates my ability to choose the optimal type of visualization to minimize viewer cognitive overload and maximize image interpretability. This is reflective of the learning outcome focused on effective communication of insights through visualization, which is crucial for influencing data-driven business decisions.

Tools:

**R:** Language used for statistical analysis and data manipulation

**RStudio:** Development environment for R

**dplyr:** Package for data transformation

**tidyr:** Package for organizing raw data

**ggplot2:** Package for creating visualizations

**Adobe Illustrator:** Software used to refine and enhance visualizations created in R